DEEP LEARNING TEXT SUMMARIZATION MODELS IN THE U.S. INTELLIGENCE COMMUNITY

by

Michael R. LeVine
Supervisory Special Agent, Diplomatic Security Service, U.S. Department of State
Student ID 51519
NIU Class 2024

Submitted to the faculty of the
National Intelligence University
in partial fulfillment of the requirements for the degree of
Master of Science and Technology Intelligence

June 2024

Deep Learning Text Summarization Models in the U.S. Intelligence Community

Thesis Accepted on Behalf of the National Intelligence University

Thesis Submitted by:

Michael R. LeVine

Michael R. LeVine

Thesis Committee:

JS Hurley Digitally signed by JS Hurley
Date: 2024.06.07 11:00:53 -04'00'

Dr. John S. Hurley, Thesis Advisor and Chair

Adam Jungdahl

Dr. Adam Jungdahl, Reader and Committee Member

ABSTRACT

TITLE OF THESIS: Deep Learning Text Summarization Models in the U.S. Intelligence Community

STUDENT: Michael R. LeVine, MSTI, 2024

CLASS NUMBER: NIU 2024 **DATE:** June 2024

THESIS COMMITTEE CHAIR: Dr. John S. Hurley

COMMITTEE MEMBER: Dr. Adam Jungdahl

Machine learning (ML) text summarization models, driven by recent advances in generative artificial intelligence (AI), efficiently reduce emails, reports, and other documents into concise summaries. Analysts in the intelligence community (IC) can benefit from these models, gaining an advantage over adversaries when time is critical. However, there is a lack of contemporary, in-depth IC research on this topic. This study addresses that gap by answering the research question: What are the most effective approaches to high-quality automatic text summarization in the IC? Experiments with pre-trained language models (BART, T5, and Pegasus) and large language models (LLMs) (Llama2 and Llama3) were conducted. Models were tested on the summarization task, and several were fine-tuned with additional training data. Summarization quality was quantitatively scored with the automated metric ROUGE. Llama3 performed qualitative comparisons between model-generated summaries. Limited human evaluation supplemented the research. The studies reveal that Llama2 appears to outperform the smaller pre-trained models in the summarization task. The research also highlights the limitations of the ROUGE metric for measuring summary quality and suggests Llama3 as an effective evaluation tool. Additionally, organizational culture and resource allocation are identified as key enablers for deploying AI/ML solutions at scale. It is hoped that this study will assist IC organizations to understand, import, test, train, deploy, and evaluate state-of-the art text summarization models in mission-critical scenarios.

ACKNOWLEDGEMENTS

I would like to thank my thesis committee chair, Dr. J.S. Hurley, for his steadfast guidance, outstanding support, utmost professionalism, and unwavering commitment to this research. I would also like to thank Dr. Adam Jungdahl, my thesis reader, for the brilliant ideas he proposed and for the deep NLP expertise and mentorship he provided which shaped the direction of this research. The thesis in its current form would not have been possible without them. I would also like to thank the incredibly supportive staff of Pacific Northwest National Laboratories, who dedicated countless hours to providing the compute infrastructure and resources necessary for this research, and for allowing me to use the high-performance computers at the iRES facility. I am deeply indebted to the science and technology group at the Johns Hopkins University Applied Physics Lab who provided guidance and input that was crucial to this research, especially Greta K. Kintzley whose professionalism and expert knowledge of LLMs was instrumental to this research. I also wish to thank the faculty of NIU, as well as colleagues from IARPA, NSA, and NIST – all of whom helped me in innumerable ways. Most importantly, I'd like to thank my loving wife, Michelle, and my two children, Jacqueline and Jonathan, whose encouragement, sacrifice, support, and laughter made this entire endeavor possible. Thank you!

CONTENTS

FIGURES

TABLES

CHAPTER 1

Introduction

Problem Statement and Relevance to the Intelligence Community

Text summarization models are machine learning algorithms that take a body of text as input and produce a summary of the input text as output.[1] Today's most performant models typically have a deep learning neural network architecture. In practical usage, text summarization models perform the function of reducing emails, reports, or other documents into concise summaries that capture the most important aspects of a source document. Analysts within the intelligence and national security communities review, distill and summarize textual information to produce finished intelligence. Therefore, text summarization models are important in the intelligence community (IC) context because they can enable faster analysis, more fulsome reporting, and can be a force multiplier by allowing a single person to review more information in less time. Crucially, these models can provide an intelligence advantage over adversaries in scenarios where rapid decisions are necessary and information is arriving faster than we can analyze. Reliable, trustworthy, and capable text summarization models would be useful for policymakers and analysts across the national security, intelligence, military, and diplomatic domains.

Organizations or teams within the IC that seek to deploy this technology at scale will need to choose a model, instantiate the model in a secure environment, and then place the model pipeline into production with sufficient compute resources to support the intended use case. At

[1] Lewis Tunstall, Leandro Von Werra, Thomas Wolf, *Natural Language processing with Transformers, Revised Edition: Building Language Applications with Hugging Face* (Sebastopol, CA: O'Reilly Media, Inc., May 2022), chapter 1, "Hello Transformers."

that point, users would be able to “run inference” on the model by providing input text from which a summary would be generated. Although a centralized, single summarization pipeline does not exist within the IC, numerous teams and organizations across the IC are experimenting with summarization models. In the commercial sector, summarization models abound: large language models like ChatGPT can perform summarization and many widely used products, like Microsoft Teams and Zoom, have built-in summarization capabilities.[2]

The IC is focused on implementing powerful, state-of-the-art AI and ML solutions because the technology can provide strategic national security advantages. Whether it is computer vision models that help satellites or intelligence, surveillance, and reconnaissance (ISR) assets detect earlier warnings of nuclear tests or launches, or natural language processing (NLP) models that enable faster and better analysis of vast corpora of signals intelligence (SIGINT), the applicability of AI solutions to the IC are numerous, especially as models become more powerful and capable. The need for AI/ML solutions aligns with U.S. Government strategy. For example, the National Security Strategy says the IC should use data tools to achieve national security objectives.[3] Furthermore, the National Intelligence Strategy includes a recommendation for the IC to use data tools that streamline “labor- and time-intensive work,” and the IC Data Strategy requires the IC to adopt AI and machine-assisted workflows.[4]

[2] “Generate a Meeting Summary,” Microsoft, February 2, 2024, https://learn.microsoft.com/en-us/microsoft-sales-copilot/generate-meeting-summary; “Enabling Meeting Summary with AI Companion,” Zoom, March 11, 2024, https://support.zoom.com/hc/en/article?id=zm_kb&sysparm_article=KB0057960.

[3] The White House, *National Security Strategy*, 46, October 2022.

[4] Office of the Director of National Intelligence, *2023 National Intelligence Strategy*, 9.; Intelligence Community Chief Data Officer, *The IC Data Strategy*: 2023-2025, 3.

Although text summarization as an area of research has been around since at least the 1970s, for decades the technology lacked the capability to be seriously considered for use in the national security domain. However, the advent of the transformer architecture in 2017 led to an almost immediate step change in the capabilities of language models and paved the way for today's highly performant generative NLP models like ChatGPT and Gemini.[5] The technology needed for high-quality text summarization models now exists. Many of today's most powerful models are open-source. With a few lines of code in a Jupyter notebook, these models can easily be brought into a compute environment from Hugging Face (Hugging Face is a leading open-source platform for machine learning models, datasets, and resources that is widely used by AI researchers and practitioners).[6] The wide availability of these models, and their power, have made it easier than ever to experiment with these models in the IC environment. Despite widespread model availability and growing importance of text summarization algorithms, there is a lack of contemporary, in-depth research on the topic within the IC.

Although the rapid technological advancements make this an important moment for the AI industry and research community, it is also a huge moment for the IC. As the nation races ahead with research and development to maintain its technological edge, it must focus on the

---

[5] Sinan Ozdemir, "A Brief History of NLP" in "Introduction to Transformer Models for NLP: Using BERT, GPT, and More to Solve Modern Natural Language Processing Tasks," (O'Reilly Media: February 2023), https://learning-oreilly-com.library.access.arlingtonva.us/course/introduction-to-transformer/9780137923717/; Aurélien Géron, "Natural Language Processing Using Transformer Architectures," presentation at Tensor Flow World, October 2019, (O'Reilly Media: February 2020), https://learning-oreilly-com.library.access.arlingtonva.us/videos/natural-language-processing/0636920373605/0636920373605-video329383/.

[6] "The AI Community Building the Future," Hugging Face, accessed February 15, 2024, https://huggingface.co/.

strategic threat that China poses in the realm of AI.[7] China recognizes that it lacks parity with the U.S. with respect to the technological capabilities of weapon systems and global power projection. Nonetheless, China believes that it can gain a military advantage by capitalizing on the ascendancy of the information domain in future conflict and has a adopted a strategy to be the leader in AI-based warfare. Chinese leaders see AI as a key component to how they will compete with the United States.[8] The IC will need unwavering dedication to and resourcing of AI and ML technologies to maintain an advantage over China, and text summarization models should be a component of this strategy. In addition to outcompeting China, the IC will also need to keep up with the rapid explosion of data volume and the velocity at which is produced, especially textual data. Sophisticated intelligence collection platforms, endless streams of SIGINT and OSINT, social media feeds, all combine to create vast amounts of intelligence that needs to be analyzed and assessed. Looking towards the future, text summarization models may eventually come to be a practical and important tool that the IC relies on for keeping up with the endless streams of text-based intelligence.

Research Overview

As organizations within the IC investigate how text summarization models can support mission needs and then make the decision to allocate resources to deploying these models at scale, they will need a clear understanding of how the models work and their limitations. Inevitably, questions will come up about optimal models, necessary compute resources, the

[7] Office of the Director of National Intelligence, *2023 National Intelligence Strategy*, 5,9.

[8] Michael Dahm, "Chinese Debates on the Military Utility of Artificial Intelligence," *War on the Rocks* (website), June 5, 2020, 1-4, 11, https://warontherocks.com/2020/06/chinese-debates-on-the-military-utility-of-artificial-intelligence/.

reliability of outputs, and how to measure performance. These are logical questions, and the purpose of this research is to answer them. This thesis examines the performance of deep learning NLP text summarization models on English language documents within the IC. Ultimately, the findings of this study can serve as a useful playbook for organizations and practitioners across the IC looking to experiment with and deploy text summarization models.

Research Question and Hypothesis

The central research question of this thesis is: What are the most effective approaches to high-quality automatic text summarization in the IC? A main hypothesis of this research is that a large language model (LLM) will outperform other models in the summarization task. The thesis also explores additional questions that help investigate the central research question. These include:

- Which text summarization models do contemporary researchers and practitioners assess to be the most performant?
- To what extent are these models modifiable, and how does that affect summary quality?
- What effect does additional training of the models, known as fine-tuning, have on summary quality?
- What level of certainty can be established about how well these models perform?
- How do summarizations generated by LLMs compare to summarizations generated by smaller models?
- How can LLMs be used to evaluate summary quality, and to what extent do the results of LLM evaluations align with automated metrics?

Research Findings

A series of experiments was conducted to address the research questions, and several important findings emerged. One finding is large language models outperform other models in their ability to summarize documents. Another finding is that the most widely used automated metric for measuring summary quality, known as the ROUGE metric, has limited usefulness and applicability to measuring the performance of today's powerful generative AI models. A corollary to this finding is that there is currently no accepted automated metric to take the place of ROUGE. This research showed that LLMs, particularly Meta's open-source model Llama3, are useful for evaluating summary quality and can partially address the gap in evaluation metrics. The research also identified critical organizational enablers that must be in place to implement text summarization pipelines in production. It is hoped that the research approach, experiments, findings, and conclusions will be practical and useful for IC organizations who are considering investing in the technology. Ideally, the information contained within this thesis will provide organizations with guidelines and processes that can be used to understand, import, test, train, deploy, and evaluate state-of-the art text summarization models.

CHAPTER 2

Literature Review

Literature Review Introduction

Automatic text summarization, described in the literature as the automated creation of text summary output from a longer text input, is an important area of study to researchers and industry.[9] The importance of the topic to the research community is demonstrated by the growth in the number of studies, papers, and books on the topic between 1980 and today. While text summarization in the field of artificial intelligence begins to appear in the literature as an area of research around 1980, and was researched dozens of times during the 1990s, the field began growing significantly during in the ten-year period from 2000 – 2009, when there were over 600 publications on the topic. From 2010 to present, there were nearly 5,000 publications on the topic, with nearly 4,000 of those occurring in the four years from 2019 to present (see figure 2.1).[10] The research, much of which appears in scholarly journals and in conferences, has focused on how model types, training datasets, model architecture, input parameters, and output parameters impact the model's summarization performance. Commercial interest in text summarization began in the 1980s, much of it within the legal field, where the technology could be used to summarize long legal documents. Today, commercialization of text summarization

---

[9] Fabian Retkowski, "The Current State of Summarization," in *Beyond Quantity: Research with Subsymbolic AI* (forthcoming, November 2023; published online August 1, 2023), 1, 6-7, https://arxiv.org/abs/2305.04853.

[10] Author's analysis of the following databases within the holdings of the National Intelligence University: ProQuest (12 databases), EBSCOHost (five databases), EBSCO Discovery, Gale Academic One File, Gale Military and Intelligence, Hein Online (over 50 databases including U.S. and military), JSTOR. Search terms were "text summarization" AND "summarization"; search timeframe from Jan 1, 1940 to October 7, 2023.

technology can be seen in products that summarize documents, conversations, meetings, messages, and sales calls.[11] Given the practical benefits of text summarization, and the ability to deploy state-of-the-art summarization models across a number of product lines, use cases, and industries, it is probable that text summarization will remain an important area of study.[12]

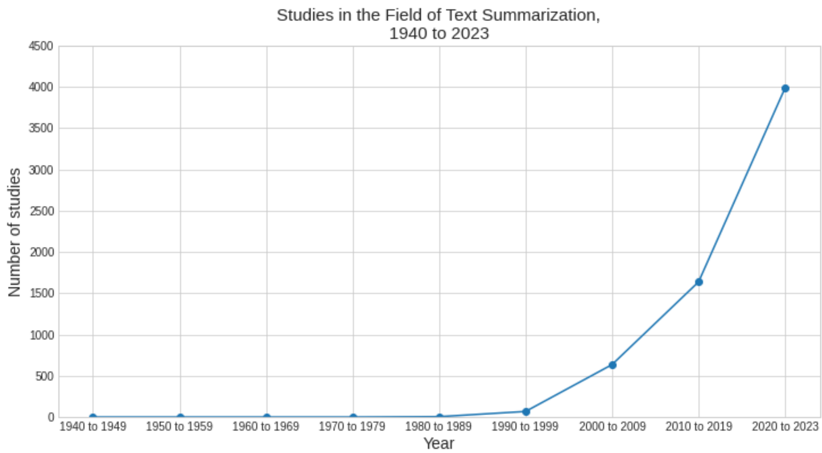

Figure 2.1 Analysis of database searches, October 2023.

The purpose of the literature review is to trace the history of scientific developments and major themes in the field of text summarization, and to describe recent trends in research methods and experiment designs. The literature review examines the characteristics of machine learning algorithms that figure prominently in the field. The metrics and measures researchers have used to evaluate the performance of summarization models, as well as current state of the

---

[11] Retkowski, "The Current State of Summarization"; Brenda Stolyar, Julian Chokkattu, "Everything Microsoft Announced at Today's Surface Event," Wired, September 21, 2023, https://www.wired.com/story/everything-microsoft-announced-surface-event-2023/.

[12] Giarelis, Mastrokostas, and Karacapilidis, "Abstractive vs. Extractive Summarization."

art models and top performers are examined in the literature review. Additionally, the literature is reviewed to identify commercialization trends and areas of the field where consensus exists and whether findings are disputed. The literature review also examines the state of research on and usage of text summarization models in the IC. The literature review begins with a condensed overview of the history of the field from 1950 to present, and then examines the period from 2019 to 2023 more closely to identify recent trends in models and performance. A corpus of more than two dozen studies, many of them peer reviewed, along with dozens of books, were used for the literature review.

History

Although natural language researchers in the 1950s were examining how computers might be used for automatic translation, and researchers had built computer systems capable of deriving limited meaning from text as early as 1964, progress was constrained by technical limitations and the systems that were developed did not perform well.[13] The earliest question-answering systems, which were designed to output answers to input questions, were unable to deduce complex meaning from the information that was stored in memory. Furthermore, a user needed to input questions in a highly structured format, and the only way these systems could answer questions was by outputting pre-existing sentences which had already been programmed

[13] Charles J. Rieger III, "Conceptual Memory: A Theory and Computer Program for Processing the Meaning Content of Natural Language Utterances," (PhD diss., Department of Computer Science, Stanford University , Apr. 1974), 1, 3-6, ProQuest Dissertations & Theses Global, https://niu.idm.oclc.org/login?url=https://www.proquest.com/dissertations-theses/conceptual-memory-theory-computer-program/docview/302726268/se-2?accountid=10504; Terry Winograd, "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language," (revised version of PhD diss., Massachusetts Institute of Technology, Jan. 1971), 229, https://dspace.mit.edu/handle/1721.1/15546.

into memory.[14] By 1968, programmers were evolving their design approach and were trying to use data structures to embed the meaning of textual information into computer systems, instead of using formal logic and syntax rules to extract meaning. The PLANNER language was developed in 1969, giving programmers an extensible set of coding expressions that they could use to input text and meaning into a computer system. However, the algorithms developed with PLANNER were still not capable of reasoning beyond the corpus of text they were programmed with.[15]

During the 1970s, researchers continued to develop novel ways to derive meaning from text, leading to the emergence of the specific field of text summarization at the end of the decade. In 1974, a system was developed that used conceptual graphs to derive meaning from text, and in 1978 researchers created a system that could infer meaning to enhance comprehension.[16] Additionally, in 1979 a system called OPUS was developed that attempted to understand references to objects in the physical world.[17] By 1980, research began to emerge in the field of automatic text summarization, with the BORIS computer system using connected graph structures and nodes to derive the conceptual meaning required to produce narrative

---

[14] Winograd, "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language," 226-244.

[15] Winograd, "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language," 240-244.

[16] Rieger III, "Conceptual Memory," 6; Richard Edward Cullingford, "Script Application: Computer Understanding of Newspaper Stories," (master's thesis, Department of Computer Science, Yale University, Jan. 1978), 7-10, in PROQUESTMS ProQuest Dissertations & Theses Global, https://niu.idm.oclc.org/login?url=https://www.proquest.com/dissertations-theses/script-application-computer-understanding/docview/288085255/se-2.

[17] Wendy Lehnert and Mark H. Burstein, "The Role of Object Primitives in Natural Language Processing," (research report #162, Department of Computer Science, Yale University, January 1979), 4, Defense Technical Information Center (DTIC), https://apps.dtic.mil/sti/citations/ADA069861.

summaries.[18] That same year, Yale University researchers published an article entitled "Narrative Text Summarization" that appeared in the First Proceedings of the American Association of Artificial Intelligence (AAAI) at Stanford University.[19] The appearance of the article at the first AAAI proceedings shows that text summarization has been considered part of the artificial intelligence field since at least 1980. Today, more than four decades later, text summarization is a well-established field of study within the AI subfield of natural language processing.[20]

Natural Language Processing (NLP) and Deep Learning Neural Networks

Today's most performant text summarization models are based on the transformer architecture which is a deep learning neural network.[21] While the use of deep learning neural networks is a common approach in NLP applications and research today, a notable development in the history NLP and deep learning occurred in 2011. In that year, a research team trained a deep learning model to recognize vocabulary words and short sentences in a large dataset of audio recordings of human speech. The team found that a key variable affecting a model's ability to recognize a string of words in the audio recordings was the depth of the neural network,

---

[18] Wendy Lehnert, "Affect Units and Narrative Summarization," (research report #179, Department of Computer Science, Yale University, January 1979), abstract (no page no.), Defense Technical Information Center (DTIC), https://apps.dtic.mil/sti/citations/ADA086735.

[19] Wendy G. Lehnert, "Narrative Text Summarization," *Proceedings of the AAAI Conference on Artificial Intelligence*, 1, (1980): 337-339, https://aaai.org/papers/00337-narrative-text-summarization/.

[20] Jyotika Singh, *Natural Language Processing in the Real World: Text Processing, Analytics, and Classification* (Boca Raton, FL: CRC Press LLC, 2023), 8, ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/niulibrary-ebooks/detail.action?docID=7250810.

[21] Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," *Applied Sciences* 13, no. 13: 7620 (June 2023): 4-5. https://doi.org/10.3390/app13137620.

meaning the number of parallel layers of neurons.[22] For example, the research showed that five-layer neural networks were 71.8% accurate at the text recognition task, four-layer networks achieved 70.2% accuracy, and three-layer networks achieved 69.6% accuracy. On the other hand, the best-performing single-layer neural networks achieved a 68.1% level of accuracy.[23] Subsequent researchers, building on the finding of the 2011 study showing the correlation between the depth of neural networks and model performance, continued to pursue research on deep learning neural networks in several fields of machine learning. In 2012, computer vision researchers used a deep neural network to classify over 1 million images in the annual image classification contest known as ImageNet. The model, called AlexNet, was the only one in that year's contest to use a deep neural network architecture. The other models competing that year all used traditional machine learning architectures, as was typically the case in previous ImageNet contest.[24] Yet the AlexNet model, which was architected with eight neural network layers, outperformed all the other entrants, winning the competition by a significant margin – its error rate was 15.3% while the next best model's error rate was 26.2%.[25] The top performing

[22] Jon Krohn, Grant Beyleveld, and Aglae Bassens, *Deep Learning Illustrated* (Boston: Addison Wesley, 2020), chapter 2.

[23] George E. Dahl et al., "Large Vocabulary Continuous Speech Recognition with Context-Dependent DBN-HMMS," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2011): 4690, https://doi.org/10.1109/ICASSP.2011.5947401.

[24] Krohn, Beyleveld, Bassens, *Deep Learning Illustrated*, chapter 2.

[25] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks" in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, ed. F. Pereira, C.J. Burges, L. Bottou and K.Q. Weinberger (Red Hook, NY: Curran Associates, Inc., 2013), 1. https://papers.nips.cc/paper_files/paper/2012.

models in the ImageNet contest over subsequent years have been deep learning models, demonstrating the capabilities of the architecture.

As computer vision researchers examined the power of deep neural networks in the years following the 2012 success of AlexNet, deep learning models were also driving advancements in the NLP field.[26] For example, in 2013 and 2014, machine translation researchers began to use neural networks for the translation task. These translation models used the encoder-decoder framework in which a neural network would read in, or "encode," a variable-length sequence of words, for example a sentence in the source language material, into a fixed-length numerical format known as a word embedding or vector space embedding. The decoder part of the model would then "decode" the fixed-length vector embedding into a variable-length output sentence, for example a translated sentence. However, an inherent limitation of the encoder-decoder model for machine translation was the need to reduce variable-length input sentences into a fixed-length vector representation during the encoding process. Long input sentences were difficult for encoder-decoder models to handle as the sentences needed to be reduced to a fixed-length during the encoding process. Research showed a marked degradation in the performance of encoder-decoder models when attempting to translate long sentences. In 2014, researchers addressed this limitation by designing an encoder-decoder machine translation model that encoded an input sentence into a succession of vector embeddings, as opposed to a single fixed-length embedding. During the encoding process, the model would link the input sequence with a series of annotations describing not only the whole sentence, but also the relationship of each word in the sentence to words that came both before and after it. The model used a bi-directional

[26] Krohn, Beyleveld, Bassens, *Deep Learning Illustrated*, chapter 2.

recurrent neural network (RNN) to examine the words before and after each word in an input sentence. A crucial feature of this model was that the decoder would rely on these annotations to decide which portions of the input sequence to pay attention to, known as an "attention mechanism." The model's novel encoding approach enabled it to produce a better translation regardless of input sentence length and addressed the limitations of previous encoder-decoder models that relied on fixed-length vectors for the encoding process. Crucially, the model devised an attention mechanism for the decoder which would lead to important developments in NLP.[27]

Having seen the success machine translation researchers had found by using deep neural networks combined with an attention mechanism, in 2015 a team of text summarization researchers explored the use of an encoder-decoder model with an attention mechanism for the summarization task. The model, which was trained on a dataset of headlines and corresponding articles, was designed to create a single sentence abstractive (i.e., generative) summary headline based upon the first sentence of a news article. The model functioned by taking an input sentence of length M and outputting a summary sentence of length N, where $N < M$. The researchers incorporated the attention-based approach to the model by including functions that calculated scores based on context words preceding a target word in the input sentence. The model was thereby using the attention mechanism to understand the meaning of a word based upon the words that preceded it. The model's architecture was similar to the 2014 attention-based model used for machine translation, although it used a feed-forward neural network as opposed to a recurrent neural network. Although the model's usefulness was somewhat limited

[27] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," conference paper at the *2015 International Conference of Learning Representations* (ICLR), (2015): 1-4, https://arxiv.org/abs/1409.0473.

in that it created a single-sentence headline by using only the first sentence of a news article, it nonetheless scored better at the summarization task than the most performant models at the time.[28] Furthermore, the research showed that deep neural networks with attention-based encoders were effective at the abstractive summarization task.

In 2017, a Google research team published a seminal paper entitled “Attention is All You Need” describing a new model architecture, the Transformer, which consequentially impacted the NLP field. When the paper was published, the best approach for conveying a sequence of input text to a desired output (the essence of NLP) was through the use of convolutional or recurrent neural networks (RNNs) based on the encoder-decoder architecture. The most performant models also added the attention mechanism. The Google team noted that a critical drawback of RNNs is that they require text sequences to be evaluated sequentially. This characteristic was preventing RNNs from performing many calculations at the same time, a capability known as parallelization which applications rely on to improve performance and speed. While the team acknowledged that RNNs with the attention mechanism were good at capturing dependencies between non-sequential words, it noted that such models were still limited by the sequential nature of RNNs, especially when evaluating longer text sequences.[29]

The Google team designed the Transformer architecture without RNNs and convolution to enable parallel processing, reduce computational complexity, and improve the model’s

---

[28] Alexander M. Rush, Sumit Chopra, and Jason Weston, “A Neural Attention Model for Abstractive Sentence Summarization,” in *Conference Proceedings: Conference on Empirical Methods in Natural Language Processing* (EMNLP), 2015, Lisbon (Red Hook, NY: Curran Associates, Inc., 2015): 1-2, 6-7 https://doi.org/10.48550/arXiv.1509.00685.

[29] Ashish Vaswani et al., “Attention is All You Need,” in *NIPS '17: Proceedings of the Annual Conference on Neural Information Processing Systems* (Red Hook, NY: Curran Associates, Inc.: December 2017):1-2, https://dl.acm.org/doi/10.5555/3295222.3295349.

training capabilities and speed. A novel aspect of the Transformer architecture was that it relied completely on the attention mechanism to understand meaning and retain semantic connection between input and output text – it did not use RNNs. Specifically, the Transformer used the self-attention mechanism to connect different parts of a text sequence and to understand the semantic dependencies between words and phrases that appear non-consecutively, especially those sequences with "long-range dependencies." The team noted that the Transformer model was the first architecture with these characteristics.[30]

The Transformer architecture used stacked encoder and decoder layers to process text. Each of the six encoder layers included a self-attention mechanism and a feed-forward, fully-connected neural network (feed-forward networks are models in which data flows in one direction, from input to output). The feed-forward neural network embeds words into unique locations within a multi-dimensional vector space to represent semantic meaning. The decoder also had six layers, each of which contained three sub-layers: a feed-forward neural network and a self-attention mechanism (just like the encoder layer), as well as a masked multi-head attention mechanism that provided an additional attention mechanism for the output sequence. By including another attention mechanism for the output sequence, which is computed one element at a time, the decoder takes the entirety of the output sequence up to that point as an additional input. Furthermore, the decoder's attention mechanism enabled the decoder to connect each element of the output sequence with all elements in the input sequence. Consequently, the attention mechanisms in the decoder improved the output sequence of the Transformer.[31]

[30] Vaswani et al., "Attention is All You Need," 2,6.

[31] Vaswani et al., "Attention is All You Need," 3,5.

The team trained two sizes of Transformers (base models with parameter sizes ranging from 28 to 168 million parameters, and a big model with 213 million parameters), and although the models were cheaper and quicker to train than other contemporary models, they nonetheless performed very well on the translation task.[32] The base and the big Transformer models outperformed all previous models on English-to-German translation, with the big model scoring 28.4 on the BLEU evaluation metric which was a new state-of-the-art score on this metric. The big Transformer model also outperformed all other previous models on the English-to-French translation task. The team also showed that with limited tuning the Transformer models outperformed almost all other previous models on another NLP task known as constituency parsing which is used to understand the grammatical relationships in a sentence.[33]

Although the Google team built the Transformer architecture for machine translation, NLP practitioners soon adopted the architecture for an array of NLP tasks because of how well these models performed.[34] Within a few months of their introduction in 2017, Transformer-based architectures were beating other models in nearly every NLP task.[35] Several prominent Transformer models were released soon thereafter, including GPT and GPT-2 in June 2018 and

---

[32] Vaswani et al., "Attention is All You Need," 8.

[33] Vaswani et al., "Attention is All You Need," 7-8.

[34] Sinan Ozdemir. "A Brief History of NLP" in *Introduction to Transformer Models for NLP: Using BERT, GPT, and More to Solve Modern Natural Language Processing Tasks*, (O'Reilly Media: February 2023), https://learning-oreilly-com.library.access.arlingtonva.us/course/introduction-to-transformer/9780137923717/.

[35] Aurélien Géron, "Natural Language Processing Using Transformer Architectures," presentation at Tensor Flow World, October 2019, (O'Reilly Media: February 2020), https://learning-oreilly-com.library.access.arlingtonva.us/videos/natural-language-processing/0636920373605/0636920373605-video329383/.

February 2019, respectively; BERT in October 2018; and BART and T5 in October 2019.[36] In fact, Transformer-based architectures are the building blocks of today's most well-known large language models, like Open AI's GPT-4 and Google's Gemini.

Text summarization researchers recognized that Transformer-based models could be effective for the summarization task, and they began to experiment with these models. By 2017, when the Transformer architecture was developed, deep learning summarization models were using convolutional neural networks (CNNs) and RNNs, as well as a variation on the RNN known as and Long Short-Term Memory (LSTM). However, research showed that Transformer-based models were better at summarization and were capable of achieving state-of-the-art results.[37] Consequently, since 2017, text summarization researchers have increasingly selected Transformer models like the popular BART (bidirectional auto-regressive transformers) and T5 (text-to-text transfer transformer) models for summarization.[38] Pegasus, as well as BRIO (which is an extension of BART) are two other Transformer-based models that are commonly used.[39] The *Papers with Code* website, which is maintained by a core team from Meta AI Research but receives contributions from across the AI/ML community, tracks the top-performing machine

[36] Abubakar Abid et al., "How Do Transformers Work" in *NLP Course*, Hugging Face, accessed January 15, 2024, https://huggingface.co/learn/nlp-course/chapter1/1

[37] Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," *Applied Sciences* 13, no. 13: 7620 (June 2023): 4-5. https://doi.org/10.3390/app13137620.

[38] Fabian Retkowski, "The Current State of Summarization," in *Beyond Quantity: Research with Subsymbolic AI* (forthcoming, November 2023; published online August 1, 2023), http://arxiv.org/abs/2305.04853.

[39] Vivek Srivasta, Savita Bhat, and Niranjan Pedanekar, "Hiding in Plain Sight: Insights into Abstractive Text Summarization," *The Fourth Workshop on Insights from Negative Results in NLP*, Association of Computational Linguistics (2023): 68, http://dx.doi.org/10.18653/v1/2023.insights-1.8.

learning models based upon contemporary research. As of February 2024, the site's abstractive summarization leaderboard showed that the best summarization models were Pegasus and variants of the BART architecture.[40] These are all Transformer models.

Text Summarization with Large Language Models

Contemporary researchers have examined how large language models (LLMs) can be used for the summarization task. LLMs, which are pre-trained language models based on the transformer architecture, are significantly larger than the best pre-trained summarization models and are better at understanding and generating natural language than smaller-scale models.[41] Language models may be compared by the number of parameters they contain. In the context of deep learning neural networks, a parameter is a modifiable value which, when taken together, allow the network to approximate functions to map an input to an output.[42] For example, in a computer vision machine learning model, the input could be an image and the output could be the model's prediction of what the image depicts – for example, a house. In summarization, the input is the source text, and the output is the summary. A model's parameter values enable it to create functions that take inputs and produce useful outputs. The BART model has more than 400 million parameters, while the Pegasus model has over 500 million.[43] In contrast, LLMs have

---

[40] "About" and "Abstractive Text Summarization," Papers with Code, accessed February 16, 2024, https://paperswithcode.com/.

[41] Shervin Minaee et al., "Large Language Models: A Survey," (Ithaca: Cornell University Library, arXiv.org, February 20, 2024): 6, https://arxiv.org/html/2402.06196v2.

[42] Laura Graesser, and Wah Loon Keng, *Foundations of Deep Reinforcement Learning: Theory and Practice in Python*, (Boston: Addison Wesley, 2020), chapter 1.5.

[43] "Pretrained Models" and "Pegasus," Hugging Face, accessed January 21, 2024, https://huggingface.co/transformers/v2.9.1/pretrained_models.html and https://huggingface.co/transformers/v2.9.1/pretrained_models.html.

many more parameters. One variation of Open AI's GPT-3 model contains 175 billion parameters and Nvidia's Megatron-Turing Natural Language Generation Model (MT-NLG) model contains over 500 billion parameters.[44] OpenAI's GPT-4 model is rumored to have more than 1 trillion parameters, but this is not confirmed.

LLMs can perform a wide range of NLP tasks, including summarization, even though they are not fine-tuned for these purposes.[45] Because of their capabilities, the NLP community has shown an increasing interest in using LLMs for summarization, especially since the release of ChatGPT in late 2022.[46] Research to ascertain how well LLMs perform summarization is active and ongoing, but the literature does not show consensus on the topic. Several studies have postulated that LLMs should perform well on the summarization task, and one team recently concluded that LLMs do in fact outperform BART and T5, both of which are fine-tuned summarization models.[47] Alternatively, LLMs have been shown to underperform fine-tuned summarization models on automated metrics while outperforming them when human evaluation is used.[48] Others have expressed uncertainties about using LLMs for the summarization task,

---

[44] Annamalai Chockalingam, Ankur Patel, Shashank Verma, Tiffany Yeung, *A Beginner's Guide to Large Language Models, Part I* (Santa Clara, CA: NVIDIA Corporation, 2023), 9.

[45] Retkowski, "The Current State of Summarization," 2-3; Sinan Ozdemir, *Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs*, (Hoboken, NJ: Addison Wesley, 2024), chapter 1.

[46] Tong Chen et al., "Enhancing Abstractive Summarization with Extracted Knowledge Graphs and Multi-Source Transformers," *Applied Sciences 13, no. 13*: 7753 (June 2023): 1. https://doi.org/10.3390/app13137753.

[47] Xiao Pu, Mingqi Gao, and Xiaojun Wan (Wangxuan Institute of Computer Technology, Peking University), "Summarization is (Almost) Dead," working paper (Ithaca: Cornell University Library, arXiv.org, September, 2023): 1-2, 4. https://doi.org/10.48550/arXiv.2309.09558.

[48] Retkowski, "The Current State of Summarization," 4.

noting concerns with hallucinations and the veracity of information contained in summaries.[49] Given the lack of consensus, more research is needed on using and evaluating LLMs for the summarization task.[50]

Abstractive and Extractive Text Summarization

Researchers have used two types of approaches for the text summarization task: extractive and abstractive summarization. Extractive summarization models select the most important sentences in the source text and then arrange these sentences into a summary.[51] Abstractive summarization models generate summaries that contain words or phrases which may not be in the source text. Researchers have presented various mathematical definitions of the text summarization task in general, as well as the abstractive and extractive approaches.[52] The authors of a 2021 paper offered this definition:[53]

---

[49] Tong Chen et al., "Enhancing Abstractive Summarization with Extracted Knowledge Graphs and Multi-Source Transformers," *Applied Sciences* 13, no. 13: 7753 (June 2023): 1, https://doi.org/10.3390/app13137753.

[50] Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," *Applied Sciences 13*, no. 13: 7620 (June 2023): 1-20, https://doi.org/10.3390/app13137620.

[51] Giarelis, Mastrokostas, and Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," 3.

[52] Tong Chen et al., "Enhancing Abstractive Summarization with Extracted Knowledge Graphs and Multi-Source Transformers," *Applied Sciences* 13, no. 13: 7753 (June 2023): 3-4, https://doi.org/10.3390/app13137753; Sheng-Luan Hou, Xi-Kun Huang, Chao-Qun Fei, Shu-Han Zhang, Yang-Yang Li, Qi-Lin Sun and Chuan-Qing Wang, "A Survey of Text Summarization Approaches Based on Deep Learning," *Journal of Computer Science and Technology* 36, no. 3 (May 31, 2021): 634, http://dx.doi.org.niu.idm.oclc.org/10.1007/s11390-020-0207-x.

[53] Sheng-Luan Hou, Xi-Kun Huang, Chao-Qun Fei, Shu-Han Zhang, Yang-Yang Li, Qi-Lin Sun and Chuan-Qing Wang, "A Survey of Text Summarization Approaches Based on Deep Learning," 2.

- The input text may be represented by X, where X contains a sequence of m words from a larger vocabulary V:

$$X = (w_1, w_2, w_3, \dots w_m), X \in V$$

- Extractive summarization produces a summary Y, where Y contains a sequence of n words, such that:

$$Y = (y_1, y_2, y_3 \dots, y_n),$$
$$n < m,\ y_i \in X$$

- Abstractive summarization produces a summary Y, where Y contains a sequence of words such that:

$$Y = (y_1, y_2, y_3 \dots, y_n),$$

$$n < m,\ y_i \in V$$

Extractive summarization has been used in the field of text summarization for decades. As early as 1958, researchers used extractive summarization based on word frequency to produce abstracts of magazine articles and technical reports.[54] From the 1990s through the early 2010s, researchers began using more sophisticated techniques for extractive summarization. These included latent semantic analysis, in which source text was modeled into matrices to derive semantic meaning, and graph theory-based approaches, in which source text sentences and their relationships were modeled into graph nodes and edges.[55]

Abstractive summarization techniques are a more recent development in the field. Around 2010, researchers began using graph-based techniques to produce abstractive summaries. By the latter half of the 2010s, researchers were experimenting with abstractive models based on

[54] H.P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development* 2, issue 2 (April, 1958): 159, https://doi.org/10.1147/rd.22.0159.

[55] Giarelis, Mastrokostas, and Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," 3-4.

deep learning neural network architectures, such as convolutional neural networks. Today's state-of-the-art abstractive text summarization models, like BART and T5, are built on the Transformer architecture.[56]

There is a consensus in the text summarization research community that abstractive summaries are superior to extractive summaries. The shortcomings of extractive summaries include the lack of a coherent flow and poor readability.[57] Researchers note that extractive summarization models produce low-quality summaries because they do not use techniques that a human would use to summarize a document. For example, extractive models do not paraphrase ideas, expand or reduce concepts when necessary, or provide context when needed.[58] Instead, extractive summaries are limited to using words and sentences that are found in the source document. On the other hand, current abstractive models use the Transformer architecture which gives them the ability to derive semantic meaning and understand the context of words and sentences in relation to other words and sentences. As a result, abstractive models produce summaries that read better and sound like a human wrote them.[59]

---

[56] Giarelis, Mastrokostas, and Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," 5.

[57] Giarelis, Mastrokostas, and Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," 1.

[58] Alexander M. Rush, Sumit Chopra, and Jason Weston, "A Neural Attention Model for Abstractive Sentence Summarization," in *Conference Proceedings: Conference on Empirical Methods in Natural Language Processing* (EMNLP), 2015, Lisbon (Red Hook, NY: Curran Associates, Inc., 2015): 1.

[59] Giarelis, Mastrokostas, and Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," 1.

Contemporary Text Summarization Research and Experiment Design

Contemporary text summarization researchers carry out quantitative research to test the performance of abstractive summarization models. The purpose of these experiments is to determine how well the models summarizes input text and the extent to which variables, such as model type and the extent of fine-tuning, impact model performance. For the experiments, researchers choose the models they want to test, and will typically select the most performant models. As of February 2024, state-of-the-art summarization models are Pegasus, mBART and BART-IT (both variations on the BART model) which are all based on the Transformer architecture.[60] To test the models, researchers use datasets that contain a corpus of source text material and associated reference summaries of the source text. Common datasets used in text summarization research are CNN/Daily Mail (a collection of news articles and human-produced "highlights"), GigaWord (a collection of headlines from international news services paired with the first sentence of the article), NYT (more than 600,000 New York Times articles with librarian-produced summaries), and WikiSum (a collection of Wikipedia articles and their associated lead sections which summarize the article), among others.[61] Source text from the dataset is input to the summarization models. In turn, the models output a summary of the input text. Metrics are used to measure the similarity of the output summary to the reference summary.

---

[60] "Abstractive Text Summarization," Papers with Code, accessed February 16, 2024, https://paperswithcode.com/.

[61] Sheng-Luan Hou, Xi-Kun Huang, Chao-Qun Fei, Shu-Han Zhang, Yang-Yang Li, Qi-Lin Sun and Chuan-Qing Wang, "A Survey of Text Summarization Approaches Based on Deep Learning," *Journal of Computer Science and Technology* 36, no. 3 (May 31, 2021): 637-639, http://dx.doi.org.niu.idm.oclc.org/10.1007/s11390-020-0207-x.

Critics have identified several concerns with text summarization datasets which are used to train and test models. One of these concerns is factual consistency. For example, reference summaries in the XSUM dataset (a collection of BBC articles and human-produced reference summaries) were shown to include information that was not found in the source text. In addition, a portion of the reference summaries in the XSUM and CNN/Daily Mail datasets were found to have other shortcomings, like presenting information in an illogical order or not including important information. The frequency with which the records in these datasets exhibited these problems ranged from 4% to over 40%.[62] The quality of datasets is important because text summarization models use these datasets for training. Researchers have expressed concern that dataset quality issues can in turn affect model performance and reliability. For instance, while the BART model is used for abstractive summarization, it was trained on the CNN/Daily Mail dataset which is characterized by extractive reference summaries. Consequently, BART tends to copy text fragments when summarizing input text, which is more characteristic of an extractive model than an abstractive model.[63]

Text Summarization Metrics and Measures

Typically, researchers use automated metrics to measure the performance of text summarization models. Frequently used metrics include ROUGE (recall-oriented understudy for

[62] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald, "On Faithfulness and Factuality in Abstractive Summarization," working paper (Ithaca: Cornell University Library, arXiv.org, May 2020): 1-4, http://arxiv.org/abs/2005.00661.

[63] Vivek Srivasta, Savita Bhat, and Niranjan Pedanekar, "Hiding in Plain Sight: Insights into Abstractive Text Summarization," *The Fourth Workshop on Insights from Negative Results in NLP, Association of Computational Linguistics* (2023): 67, http://dx.doi.org/10.18653/v1/2023.insights-1.8.

gisting evaluation), BLEU (bilingual evaluation understudy), BERTScore, and BARTScore.[64] Several variations of the aforementioned metrics are ROUGE-n (where n equals a number such as 1, 2, 3, etc.), ROUGE-L, ROUGE-SU, BLEU-1, BLEU-2 and SACREBLEUE.[65] While the aim of all of these metrics is to measure the quality of summary, they differ somewhat in approach. ROUGE scores are computed by quantifying how many n-grams in the reference summary are found in the model's output summary. An n-gram is a sequence of words, where n is the length of a sequence. ROUGE-1 measures unigrams, or single words; ROUGE-2 measures bi-grams, or two-word sequences; and ROUGE-L measures the longest sequence of words from the reference summary that is found in the model summary.[66] ROUGE scores range from zero to one, where a lower value indicates a lack of similarity between the output summary and the reference summary, and a score closer to one indicates more similarity between the two.[67] Like ROUGE, the BLEU family of metrics also measures overlapping n-grams, or sequences of words. However, whereas ROUGE measures the number of n-grams from the

[64] Fabian Retkowski, "The Current State of Summarization," in Beyond Quantity: Research with Subsymbolic AI (forthcoming, November 2023; published online August 1, 2023): 3-4, http://arxiv.org/abs/2305.04853; Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," Applied Sciences 13, no. 13: 7620 (June 2023): 7, https://doi.org/10.3390/app13137620.

[65] Janani Ravi, "Evaluation Metrics for Summaries" in "AI Text Summarization with Hugging Face," October 30, 2023, LinkedIn Learning, https://www.linkedin.com/learning/ai-text-summarization-with-hugging-face/evaluation-metrics-for-summaries; Giarelis, Mastrokostas, and Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," 7.

[66] Alexander R. Fabbri et al., "SummEval: Re-Evaluating Summarization Evaluation," *Transactions of the Association for Computational Linguistics* 9 (2021): 393, https://doi.org/10.1162/tacl_a_00373; Chris Fregly, Antje Barth, Shelbee Eigenbrode, *Generative AI on AWS: Building Context-Aware Multimodal Reasoning Applications* (Sebastopol, CA: O'Reilly Media, Inc., December, 2023), chapter 5 "Fine-Tuning and Evaluation."

[67] Janani Ravi, "Evaluation Metrics for Summaries" in "AI Text Summarization with Hugging Face," October 30, 2023, LinkedIn Learning, https://www.linkedin.com/learning/ai-text-summarization-with-hugging-face/evaluation-metrics-for-summaries.

reference summary that are found in the model-generated summary, BLEU measures the number of n-grams in the generated summary that are also found in the reference summary. BLEU scores also use a zero to one scale, with lower scores indicating a lack of similarity between the generated summary and the reference summary, and higher scores indicating more similarity.[68]

Although metrics are widely used in the text summarization field to measure summary quality, researchers have raised concerns about the validity and reliability of these measures and have noted that better metrics are needed.[69] For instance, even though ROUGE is the most widely used summarization metric in the field, researchers acknowledge it is a flawed measure.[70] ROUGE is criticized for failing to measure how well a summary communicates key information and for not capturing the extent to which a generated summary accords with facts in the source text.[71] Studies showed ROUGE to be a problematic metric when evaluating summaries of scientific works and meetings.[72] Additionally, a research team found that ROUGE gave high

---

[68] Lewis Tunstall, Leandro Von Werra, Thomas Wolf, *Natural Language processing with Transformers, Revised Edition: Building Language Applications with Hugging Face* (Sebastopol, CA: O'Reilly Media, Inc., May 2022), chapter 6, “Summarization.”

[69] Retkowski, “The Current State of Summarization,” 6; Tunstall, Von Werra, Wolf, *Natural Language processing with Transformers, Revised Edition,* chapter 6, “Summarization.”

[70] Tong Chen et al., “Enhancing Abstractive Summarization with Extracted Knowledge Graphs and Multi-Source Transformers,” *Applied Sciences* 13, no. 13 (June 2023): 10, https://doi.org/10.3390/app13137753.; Retkowski, “The Current State of Summarization,” 4.

[71] Vivek Srivasta, Savita Bhat, and Niranjan Pedanekar, “Hiding in Plain Sight: Insights into Abstractive Text Summarization,” *The Fourth Workshop on Insights from Negative Results in NLP, Association of Computational Linguistics* (2023): 68, http://dx.doi.org/10.18653/v1/2023.insights-1.8.

[72] Alexander R. Fabbri et al., “SummEval: Re-Evaluating Summarization Evaluation,” *Transactions of the Association for Computational Linguistics* 9 (2021): 392, https://doi.org/10.1162/tacl_a_00373.

scores to summaries that were missing important information or contained hallucinations.[73] Studies also identify weaknesses of BLEU for the evaluation of model-generated summaries. Since BLEU, in its most basic form, tallies the number of words in a generated summary that appear in a reference summary, the measure can be misleading. For example, a generated summary containing a single, repeated word could theoretically score very high on BLEU provided the word appeared at least once in the reference summary. Although this example is exaggerated, and variations of BLEU guard against such simple manipulations, BLEU is nonetheless problematic because it doesn't consider synonyms and fails to capture semantic meaning.[74] Contemporary researchers have shown that evaluation metrics for generated summaries do not correspond well with human assessments and have further pointed out the challenges of comparing evaluation protocols across different summarization studies.[75] It is therefore logical to question why these metrics are still used. One reason is that ROUGE is a long-standing and familiar measure in the text summarization field that was introduced in 2004.[76] ROUGE is easy to deploy which makes the interpretation and comparison of ROUGE scores relatively straightforward.[77] Researchers have simply not been able to develop better

[73] Srivasta, Bhat, Pedanekar, "Hiding in Plain Sight: Insights into Abstractive Text Summarization," 70.

[74] Tunstall, Von Werra, Wolf, *Natural Language processing with Transformers, Revised Edition: Building Language Applications with Hugging Face*, chapter 6, "Summarization."

[75] Retkowski, "The Current State of Summarization," 4.

[76] Chin-Yew Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out* (Barcelona, Spain: Association for Computational Linguistics, 2004): 1, https://aclanthology.org/W04-1013.

[77] Srivasta, Bhat, Pedanekar, "Hiding in Plain Sight: Insights into Abstractive Text Summarization," 68-69.

alternatives to the existing metrics.[78] Even though they are still used, the literature shows consensus on the need for new evaluation metrics and measures to address the inadequacy of the current metrics, and this is an area of active research.[79]

Human evaluation is an established evaluation method in the text summarization field, and contemporary researchers augment their studies with human evaluation to complement automated metrics. Human evaluation of summaries had been in use for decades before the introduction of automated metrics around 2000.[80] Human evaluators judge generated summaries on several dimensions like how well the parts of a summary connect to each other, whether the summary reads well, and the extent to which important information in the source text is captured in the summary. However, human evaluation is costly and difficult to scale to larger inquiries.[81] Studies have also reported variations in how evaluators interpret the criteria for scoring a summary which can result in a lack of consistency across human evaluations.[82] Despite these issues, contemporary researchers employ human evaluation to enhance the validity of studies

[78] Alexander R. Fabbri et al., "SummEval: Re-Evaluating Summarization Evaluation," *Transactions of the Association for Computational Linguistics* 9 (2021): 392. https://arxiv.org/abs/2007.12626.

[79] Srivasta, Bhat, Pedanekar, "Hiding in Plain Sight: Insights into Abstractive Text Summarization," 70; Fabbri et al., "SummEval: Re-Evaluating Summarization Evaluation," 393; Tunstall, Von Werra, Wolf, *Natural Language processing with Transformers, Revised Edition: Building Language Applications with Hugging Face*, chapter 6, "Summarization."

[80] Inderjeet Mani, *Automatic Summarization* (Philadelphia: John Benjamins Publishing Company, 2001), 221-229.

[81] Chin-Yew Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," 1.

[82] Mani, *Automatic Summarization*, 225.

and to avoid overreliance on automated metrics.[83] Research has in fact called for future studies to use human evaluation as a way to address the untoward characteristics of automated metrics.[84]

Over the last two years, researchers have developed novel approaches to measuring summarization quality, and to the summarization task itself. For example, a 2023 study divided summarization quality into three dimensions: "information coverage," "entity hallucination," and "summarization complexity." The team used a several metrics and algorithms to measure summarization quality along these dimensions, including the C99 algorithm to measure the similarity of sentences, the YAKE method to extract keywords, and a named entity recognition models to identify hallucinations. The researchers concluded that relying on ROUGE benchmarks is not an adequate approach.[85] A separate 2023 study introduced a consolidated evaluation toolkit that consolidated a range of automated metrics and other resources to assist with the evaluation of summarization models. While novel, the study did not propose any metrics to replace ROUGE or BLEU.[86] Researchers have also used LLMs to not only evaluate summaries, but to perform the summarization task as well. 2022 and 2023 studies concluded that LLMs perform well as evaluators on NLP tasks, and one of these found that LLMs align more closely with human evaluation than any automated metrics. A 2023 study concluded that

[83] Fabbri et al., "SummEval: Re-Evaluating Summarization Evaluation," 391-393.

[84] Retkowski, "The Current State of Summarization," 7.

[85] Vivek Srivasta, Savita Bhat, and Niranjan Pedanekar, "Hiding in Plain Sight: Insights into Abstractive Text Summarization," *The Fourth Workshop on Insights from Negative Results in NLP, Association of Computational Linguistics* (2023): 68-70, http://dx.doi.org/10.18653/v1/2023.insights-1.8.

[86] Fabbri et al., "SummEval: Re-Evaluating Summarization Evaluation," 401.

LLMs produce summaries of news articles that are as good as summaries written by humans.[87] Other studies, however, identified concerns with using LLMs for the summarization task, citing issues relating to hallucinations and summarization accuracy.[88] While some of these developments seem promising, the text summarization community has not reached a consensus on new approaches to measuring summarization quality and how to use LLMs for the task. Therefore, the use of established metrics coupled with human evaluation is still an accepted and common method in the field.

Text Summarization Research Gaps and Opportunities

Despite the breadth of contemporary research into text summarization and widespread interest in NLP, especially since the 2022 introduction of ChatGPT, the literature review identified gaps in text summarization research. For the national security enterprise, a notable gap is the limited amount of contemporary, published research on the use of text summarization systems within the IC. Although practitioners within the IC are currently examining and experimenting with summarization systems, and older studies exist, only three contemporary research papers on summarization within the IC were identified. The National Intelligence University's thesis database, a repository that houses IC research from 1976 to present, was searched for the terms "natural language processing" and, separately, "summarization," and the results were analyzed. Of the three studies that mentioned summarization, a 2022 study that examined a range of NLP techniques to improve intelligence analysis in the Coast Guard was the

[87] Tianyi Zhang et al., "Benchmarking Large Language Models for News Summarization," *Transactions of the Association for Computational Linguistics* 12 (2024): 39-57, https://doi.org/10.1162/tacl_a_00632.

[88] Retkowski, "The Current State of Summarization," 4-5.

only recent NIU research that treated summarization in any depth. For that study, the researcher did not use automated metrics like ROUGE or BLEU to evaluate the model-produced summaries, and in fact recommended that future studies of summarization employ these tools.[89] Additionally, two 2021 research papers funded under IARPA's Machine Translation for English Retrieval of Information in Any Language (MATERIAL) program were identified that had a nexus to summarization. One used ML models to summarize machine-translated text. The study concluded that more research on abstractive summarization is needed to make these models more useful for users.[90] The other IARPA study looked only at extractive approaches to summarization.[91] Other summarization studies identified in the DTIC search were either very dated (such as 1990s and early 2000s) or they did not treat text summarization in any depth.[92] Consequently, this thesis finds that deep learning abstractive text summarization models have not been well-studied in the IC. Therefore, this thesis aims to address a knowledge gap in the IC by providing a current analysis of an important NLP field that is changing quickly.

---

[89] Tatiana Torres, "Natural Language Processing to Improve U.S. Coast guard Intelligence Analysis and Reporting," (master's thesis, National Intelligence University, June 2022), 65, National Intelligence University thesis database.

[90] Kathleen McKeown et al, "System for Cross-Language Information Processing, Translation and Summarization (SCRIPTS)," (report number AFRL-RH-WP-TR-2021-0113, Air Force Research Laboratory, December 2021), 125, Defense Technical Information Center (DTIC), https://apps.dtic.mil/sti/trecms/pdf/AD1165721.pdf.

[91] John Makhoul et al., "Foreign Language Automated Information Retrieval (FLAIR)/Machine Translation for English Retrieval Of Information In Any Language (MATERIAL)," (report number AFRL-RH-WP-TR-2021-0088, Air Force Research Laboratory December 2021), Defense Technical Information Center (DTIC), https://apps.dtic.mil/sti/trecms/pdf/AD1165370.pdf.

[92] Author's search of DTIC database.

Another research gap identified during the literature review was the limited coverage of how LLMs fit into the summarization field, specifically with respect to their usage for evaluating summary quality. This thesis covers the use of LLMs both to generate summaries and to evaluate their quality.

Multi-document summarization, or the creation of summaries from more than one input document, is also an area that is not well-studied in the body of text summarization research.[93] While multi-document summarization is an area that warrants further inquiry, it is not within the scope of this thesis.

Literature Review Conclusion and Scoping Statement

The historical and contemporary literature on text summarization was reviewed. Key findings from the literature review informed the approach and scope of the thesis as follows:

- **Abstractive text summarization is the focus of the study**. Abstractive models are *generative* – they create new words and phrases to summarize input text. In contrast, extractive summaries rearrange words, phrases, and sentences from the input sequence to create a summary. The literature review showed that contemporary text summarization researchers find abstractive summaries to be easier to read and better than extractive summaries. Therefore, this study evaluates the most performant state-of-the-art abstractive summarization models because they perform better and are more likely to benefit the IC.

[93] Retkowski, "The Current State of Summarization," 7.

- **The study evaluates the performance of summarization models that use Transformer-based architectures**. The literature review demonstrated that summarization models which use the Transformer architecture are the most performant and are considered state-of-the-art. The Transformer's attention mechanism is the key technological advancement that makes them so good at NLP tasks. This study tests BART, T5, Pegasus, all of which use the Transformer architecture. The study also tests the summarization capabilities of Llama2 and Llama3 LLMs which are also based on the Transformer architecture.
- **This study focuses on the performance of summarization models within the national security and IC domain**. A review of the literature showed that although text summarization models are very well-studied in academia and industry, there are only a handful of recent studies of the topic within the IC. Because text summarization has the potential to benefit a wide array of use individuals across the IC, including analysts, policymakers, and planners (or anyone who needs to review a lot of textual information in a short time), it is important to understand how these models work within the IC. As such, this study examines the performance of text summarization models within the IC and national security domain.
- **This study uses automated metrics complemented by LLM evaluation to measure the quality of summaries**. The literature review showed consensus across the research community that automated metrics, like ROUGE and BLEU, are less than ideal for evaluating the quality of model-produced summaries. These metrics have been shown to be problematic for determining how well a summary communicates key information,

whether it is factual, whether key information is omitted, and the extent to which semantic meaning is captured. Since NLP researchers have not been able to develop better metrics, contemporary text summarization researcher design still uses ROUGE and BLEU, but also uses human evaluation to augment the findings and reduce the uncertainty inherent in these metrics. The present study will use the ROUGE metric and will augment the findings with LLM evaluation of summarization quality.

- **Although the research portion of this study concluded on March 1, 2024, additional advancements in the text summarization field will occur**. The technology underpinning deep-learning NLP models changes quickly, and the field is actively researched. The models, methodologies, and approaches to the summarization problem will continue to evolve rapidly. The information in this thesis, such as top models and common research approaches, is relevant and accurate as of March 1, 2024. Technological developments after this date will not be addressed in this work because of time constraints related to the NIU program.
- **This research does not cover, or covers in a limited way, certain topics**. This thesis has a limited scope and certain areas of text summarization are not covered, including the following: multi-document summarization; summarization of non-English documents; summarization across different modes (such as audio and video). Similarly, certain other areas of text summarization are covered in only a limited way due to time constraints, including the extent to which summarization models hallucinate.

CHAPTER 3

Research Methodology and Experiment Design

Research Methodology

The research used a mixed methods research approach. Qualitative techniques were applied to examine how summarization models work. A quantitative approach was used to examine the performance metrics and measures of the summarization models, and to identify which model variables and parameters impacted model performance. The mixed methods approach was appropriate because it allowed for the exploration of two key components of the research. First, explaining how summarization models work was best achieved with qualitative research. Second, testing the sensitivity of variables on measures of summary quality was best achieved with quantitative research.

To evaluate the performance of text summarization models, the research used an experimental design. This was the most relevant method for testing the performance of text summarization models because it enabled the testing of how variables like model type and parameters impacted summarization quality metrics. Controls were established for the experiments to increase the level of certainty of findings and to allow for reproducible results.

Assumptions

The following assumptions underpinned the research:

- Different models produce summaries that differ from each other and are of varying quality.

- It is not possible to determine how or why a model includes information in a summary.

- The state-of-the-art in text summarization and NLP will continue to advance rapidly.

Overview of Experiments

The research question was addressed with a series of experiments. The results of these experiments were analyzed quantitatively using statistics, graphs, and data visualizations. Additionally, the results were analyzed qualitatively to inform findings and conclusions. The following experiments were conducted:

- **Zero-shot summarization with pre-trained models:** BART-base (BART), T5-small (T5), and Pegasus-large (Pegasus) were tested in a zero-shot scenario with no finetuning. The models generated summaries of 200 CNN/Daily Mail news articles (the test set) without having been fine-tuned on a training dataset from the same corpus. The models were deployed in what is referred to as an "off-the-shelf" configuration meaning the model hyperparameters were kept at default settings. The summaries that the models generated were then compared to the reference summary and scored using the ROUGE metric. The purpose of this experiment was to evaluate how well the pre-trained models generalize to a summarization task on new data.
- **Fine-tuned summarization with pre-trained models:** Fine-tuning is the process of using a training dataset to adjust a model's pre-trained parameters so that it can perform a specific downstream task better. BART and T5 were fine-tuned using 2,871 CNN/Daily Mail article/summary pairs. The fine-tuned models were then used to summarize the 200-article test set. The summaries that the models generated were then compared to the reference summaries and scored using the ROUGE metric. The ROUGE scores were compared to the original zero-shot scores. The purpose of this experiment was to

perform supervised training using a corpus of task-specific, labeled training data to determine how the additional training affected summarization performance. Pegasus, the largest of the three pre-trained models at 568 million parameters, was not fine-tuned.

- **Hyperparameter testing of T5:** A hyperparameter sweep of three model training hyperparameters (learning rate, training epochs, and batch size) was conducted during T5 fine-tuning. In machine learning, hyperparameters are adjustable configurations that set the conditions for model training. The purpose of these experiments was to determine the extent to which modifying hyperparameters during fine-tuning affects summarization performance. Summarization performance was tested after making each of the following adjustments in the model training loop: learning rate was set to .1, .01, and .001; training epochs were set to 2, 4, and 8; batch size was set to 8, 16 and 32. During hyperparameter testing, all hyperparameters were held constant as control variables except the hyperparameter being tested, which became the independent variable (the quality of summarization as measured by ROUGE was the dependent variable). After each hyperparameter modification, T5 was used to generate summaries for the 200-article test set. The summaries that the model generated were then compared to the reference summaries and scored using the ROUGE metric.
- **Large language model (LLM) summarization using Llama-2 70 billion parameter model:** Meta's Llama-2 70 billion parameter LLM was used to generate summaries of the 200-article test set. The purpose of this experiment was to evaluate the summarization capabilities of an LLM and compare this performance to the smaller language models (BART, T5, and Pegasus). Llama-2 was prompted as follows: "Your

task is to summarize articles. You will be given an article and must produce a summary of the article." The summaries that Llama2 generated were then compared to the reference summaries and scored using the ROUGE metric.

- **Qualitative evaluation of summaries using the LLama-3 70 billion parameter model:** Meta's LLama-3 70 billion model was used as an evaluator of summary quality. The model was provided with a news article and a series of model-generated summaries and was instructed to select the best summary. The model was instructed as follows: "You are an expert evaluator for judging summaries. You will be given an article and [a set of] different summaries. Your task is to choose which summary is the best." The purpose of this experiment was to supplement the ROUGE scores with a qualitative evaluation that could, to an extent, emulate human evaluation. The Llama-3 model was used to compare summaries across the following dimensions for each record in the 200-article test set.

Experiment	Description
Off-the-shelf comparison: BART vs. T5 vs. Pegasus	The LLM was given the summaries that BART, T5, and Pegasus generated and was instructed to select the best one. These were the summaries produced by the models in an off-the-shelf configuration (i.e. no fine-tuning).
Fine-tuned comparison: T5 vs. Bart	The LLM was given the summaries generated by the T5 fine-tuned and BART fine-tuned models and was instructed to select the best one.
T5 comparison: T5 vs. T5 fine-tuned	The LLM was given the summaries generated by the T5 model and the T5 fine-tuned model and was instructed to select the best one.
BART comparison: BART vs. BART fine-tuned	The LLM was given the summaries generated by the BART model and the BART fine-tuned model and was instructed to select the best one.
All summaries	The LLM was given summaries generated by the following six models and was instructed to choose the best one: BART, BART fine-tuned, T5, T5 fine-tuned, Pegasus, Llama-2 70B.

Table 3.2 Overview of Llama3 (LLM) Experiments to Test Summarization Quality.

The experimental approach was chosen to enable an analysis of model performance along several dimensions. Testing state-of-the art pre-trained language models (BART, T5, and Pegasus) in both off-the-shelf and fine-tuned configurations showed how the models compare to each other the effect fine-tuning had on summarization performance. Experiments with Llama-2 70B highlighted its summarization abilities compared to smaller language models. The hyperparameter experiments showed the extent to which hyperparameter changes during model training affect downstream summarization tasks. Experiments with Llama-3 70B as an expert evaluator augmented ROUGE metrics with ground truth testing to understand how ROUGE metrics compare to qualitative assessments. Using an LLM as an evaluator also revealed insights about the effectiveness of using LLMs to replicate human evaluation in the text summarization field.

Compute Environment

The baseline compute environment used for the experimentation was a GPU-enabled web-based Jupyter Lab notebook environment operated and maintained by the Pacific Northwest National Laboratory (PNNL), a Federally Funded Research and Development Center (FFRDC) operated by Battelle for the Department of Energy. The notebooks used for the experiments were running Python version 3.11.5. PyTorch-CUDA-12.2 was used to train deep learning models in a GPU environment. The baseline compute environment was running a single NVIDIA A10 GPU with 24GB of GPU RAM. The fine-tuning of the BART and T5 models and the hyperparameter testing was conducted on a high-performance computer located at the PNNL iRES facility in Washington, D.C. The experiments at the iRES facility used the same compute environment consisting of the Jupyter Lab notebook environment running the same Python and

PyTorch-CUDA versions. However, the iRES workstation had better GPU capabilities – it was a dual GPU system with 96GB of GPU RAM. The Llama2 and Llama3 testing was conducted at the Johns Hopkins University Applied Physics Lab (APL), another FFRDC, in Laurel, MD. The compute environment was APL's internal, unclassified compute environment.

Dataset

The initial intent of the research was to utilize a dataset of U.S. Department of State cables, which included well-crafted human summaries that could have served as valuable reference summaries. However, this dataset could not be utilized due to the absence of a data sharing agreement that would have permitted the transfer of sensitive cables from the State Department environment to the compute environment used in this study. The dataset used in the experiments was the unclassified CNN/Daily Mail dataset. The dataset was selected for its accessibility, large number of article/summary pairs, and widespread use in text summarization research. Although the CNN/Daily Mail dataset was a reasonable option for the research that produced useful findings, experimenting with the State Department dataset could have led to insights into critical areas relevant to the IC. For example, the State Department dataset may have helped address the question of how well the language models generalize to corpora of text in the national security domain and the extent to which additional model training and fine-tuning improve summarization performance.

In addition to the CNN/Daily Mail dataset, several other sources of data were used. One of these was the results of the text summarization experiments, which were collected and analyzed after the experiments were conducted. The research also drew upon data and insights from consultations with intelligence community professionals and subject matter experts regarding the current utilization of text summarization models. Consultations were used for guidance and background purposes and the information developed during consultations is not used as evidence in the research. No additional human sources were involved in this research.

Dataset Overview

The CNN/Daily Mail dataset consists of 311,971 records, with each record consisting of a news article and a summary. The articles were written by journalists from CNN and the Daily Mail and the summaries are human-generated. The dataset's human-produced summaries were used as reference summaries when testing model performance. The reference summaries were compared to the summaries the models generated to evaluate the quality of the model's output.

Dataset Capture/Dataset Reduction

Training data was acquired through Hugging Face, which maintains a repository of nearly 150,000 ML/AI datasets at huggingface.co/datasets. The dataset was extracted from Hugging Face by importing the `'datasets'` and `'load_datasets'` libraries into a Jupyter Lab environment and then using a loading script to download the CNN/Daily Mail dataset from Hugging Face. The CNN Daily Mail was in the `'DatasetDict'` format, which is a dictionary that contains Hugging Face dataset objects. The features in the dataset objects were `['article', 'highlights', 'id']`, meaning that each record contained an article, highlights (i.e. human-generated summary), and a unique ID number.

The full CNN/Daily Mail dataset of over 300,000 records was too large for the available compute resources. A subset of the dataset was used. A sample of 2,871 records, less than 1% of the entire dataset, was taken to be used as a training set. A separate sample of 200 other records was taken from the full dataset to be kept as a holdout, or test dataset. The holdout set was labeled data that was not used during for model training. It served as a final evaluation set to assess how well the trained models generalized to unseen data which is a standard approach for testing ML models.

Dataset Characteristics

The Hugging Face dataset card for the CNN/Daily Mail dataset shows that for all the 300k+ articles in the dataset, the mean token count for the articles is 781 and the mean token count for the highlights is 56.[94] Token count does not equal word count. A somewhat useful rule-of-thumb is that one token equals about .75 words. For the 200-record holdout dataset, descriptive characteristics of the word counts were calculated and are shown below:

Object	**Minimum word count**	**Maximum word count**	**Mean word count**
Article	144	1828	735
Highlights (summaries)	13	154	53

Table 3.2 Descriptive characteristics of training dataset.

Data Preprocessing and Data Cleaning

The dataset as extracted from Hugging Face was preprocessed before being input into the BART, T5 and Pegasus models. Characters that do not impart meaning like new lines, backslashes, and double-hyphens were removed. Capitalized letters were changed to lowercase which allowed the use of uncased models and tokenizers that are less compute-intensive. The data cleaning was accomplished by creating a `clean_text` function that iterated over the text

[94] "Datasets: cnn_dailymail" in *Hugging Face/datasets*, Hugging Face, accessed May 8, 2024, https://huggingface.co/datasets/cnn_dailymail.

and replaced various strings with the empty string. The function was then mapped to the entire training and test datasets using the `.map()` method.

Tokenization was another necessary pre-processing for the BART, T5, and Pegasus models. A tokenizer breaks text into smaller units called tokens based on a predefined set of rules. Tokens are then transformed into numerical representations and converted into unique tensors that are input to a model. The tokenizer also adds any additional necessary inputs for the model, such as end of string tokens, classifier tokens, and attention masks.[95]

Models

The following language models were used for the experiments:

- **T5-small –** The text-to-text transfer transformer (T5) was selected for the summarization experiments because contemporary researchers consider it to be a performant, state-of-the-art transformer model for summarization.[96] The T5 is a sequence-to-sequence language model that uses the encoder and decoder of the transformer architecture.[97] The T5-small variant has 60 million parameters.[98] This variant was selected because it was small enough to run inference on within the PNNL compute environment (single

---

[95] "Preprocess" in *Hugging Face/transformers*, Hugging Face, accessed May 8, 2024, https://huggingface.co/docs/transformers/v4.38.2/en/preprocessing.

[96] Fabian Retkowski, "The Current State of Summarization," in *Beyond Quantity: Research with Subsymbolic AI* (forthcoming, November 2023; published online August 1, 2023), http://arxiv.org/abs/2305.04853.

[97] Sinan Ozdemir, "Encoders and decoders welcome: T5's architecture" in *Introduction to Transformer Models for NLP: Using BERT, GPT, and More to Solve Modern Natural Language Processing Tasks*, (O'Reilly Media: February 2023), https://learning-oreilly-com.library.access.arlingtonva.us/videos/introduction-to-transformer/9780137923717/9780137923717-TMN1_01_10_01/.

[98] "Model Card for T5 Small," Hugging Face, accessed May 10, 2024, https://huggingface.co/google-t5/t5-small.

NVIDIA A10 GPU with 24GB of GPU RAM) and small enough to train/fine-tune on the iRES compute environment (dual GPU system with 96GB of GPU RAM). The T5-small model was downloaded from Hugging Face and brought into the PNNL and iRES compute environment.

- **BART-base –** The bidirectional and autoregressive transformer (BART) model was selected for the summarization experiments because it is recognized as a state-of-the-art model for abstractive text summarization.[99] BART is a sequence-to-sequence transformer model that uses a bidirectional encoder and a decoder.[100] The BART-base variant has 139 million parameters.[101] This variant was selected because it was small enough to run inference on and fine-tune within the iRES compute environment (dual GPU system with 96GB of GPU RAM). The BART-base model was downloaded from Hugging Face and brought into the iRES compute environment.
- **Pegasus-large –** The pre-training with extracted gap-sentences for abstractive summarization (Pegasus) model was selected for the summarization experiments because it is recognized as a state-of-the-art model for abstractive summarization.[102] Pegasus is sequence-to-sequence transformer model that, like BART, uses a bidirectional encoder

[99] "Abstractive Text Summarization," Papers with Code, accessed February 16, 2024, https://paperswithcode.com/.

[100] "Model Card for BART (base-sized model)," Hugging Face, accessed May 10, 2024, https://huggingface.co/facebook/bart-base.

[101] Nadira Povey, "BART Model Architecture," Medium, December 12, 2022, https://medium.com/@nadirapovey/bart-model-architecture-8ac1cea0e877.

[102] "Abstractive Text Summarization," Papers with Code, accessed February 16, 2024, https://paperswithcode.com/.

and decoder. The Pegasus-large variant has 568 million parameters.[103] This variant was selected because it was small enough to run inference on within the iRES compute environment (dual GPU system with 96GB of GPU RAM). Fine-tuning was not possible given the model's size. The Pegasus-large model was downloaded from Hugging Face and brought into the iRES compute environment.

- **Llama-2 and Llama-3 –** Meta's Llama-2 70 billion parameter model and Llama-3 70 billion parameter model were selected for the summarization experiments because these were the LLMs that were available within the Johns Hopkins University Applied Physics Lab (APL) compute environment in Laurel, MD. Llama-2 was used to summarize 200 CNN/Daily Mail articles, and Llama-3 was used to perform qualitative evaluation of article summaries produced by all the other language models. The Llama-2 and Llama-3 experiments were conducted on-site at APL using APL's internal, unclassified network. The Llama models had previously been instantiated within APL's compute environment.

Test Conditions

Dependent and independent variables

The theoretical framework that underpinned the experiments presupposed a causal relationship between the models, their parameters, and the generated summary. Throughout the experiments, a host of variables were examined. In each test, the dependent variable was the model-generated summary, which was measured either quantitatively using the ROUGE metric, or qualitatively using the LLM as an evaluator. The model-generated summary was selected as

[103] "Pegasus," Hugging Face, accessed May 10, 2024, https://huggingface.co/docs/transformers/main/model_doc/pegasus.

the dependent variable because it was the variable that, when measured, could best address the research question. Another reason why the model-generated summary was selected as the dependent variable is because of the potential for it to be affected by changes to other variables during the experiment. By selecting the model-generated summary as the dependent variable, the experiment sought to investigate whether a specific treatment to other test conditions and variables influenced the quality of the model-generated summary, and how sensitive the dependent variable was to changes in the experimental conditions. The ultimate objective for selecting the dependent variable was to establish which factors predict or influence summary quality.

The experiments identified and varied several independent variables during the experiment to determine how the variation affected the dependent variable. One of the most important independent variables was the model itself (T5, BART, Pegasus, and Llama-2) because a fundamental goal of the experiments was to determine the effect that model choice has on summary quality (tokenizer selection is typically tied to model selection). Another independent variable was whether the models underwent fine-tuning. This independent variable was useful for assessing how fine-tuning affects summarization quality. During hyperparameter testing, several hyperparameters were selected, one at a time, to be the independent variable. These were learning rate, batch size, and training epochs. These independent variables were used to measure the extent to which changes in hyperparameters affected summarization quality.

During testing, only a single independent variable was changed at a time, while other control variables were held constant. This was done to isolate and study the effects that changes to a single independent variable had on the dependent variable (summary quality) and to eliminate the possibility that other factors were affecting the outcome of the experiments. Examples of control variables that were held constant during the experiments were: training and test datasets; compute environments; metrics; data cleaning; and data preprocessing. During hyperparameter testing, all hyperparameters were held constant except for the hyperparameter being tested. Examples of the hyperparameters that were held constant were: evaluation strategy (epochs), weight decay, compute metrics, and 16 vs. 32 bit training precision.

CHAPTER 4

Results

Summarization Results

T5, BART, Pegasus and Llama-2 were used to generate summaries of the 200 articles within the test dataset. T5 and BART were fine-tuned on the training set of 2,871 article/summary pairs and then used to summarize the 200 articles in the test dataset again. The below visualization shows the word count of the summaries that each of the models produced:

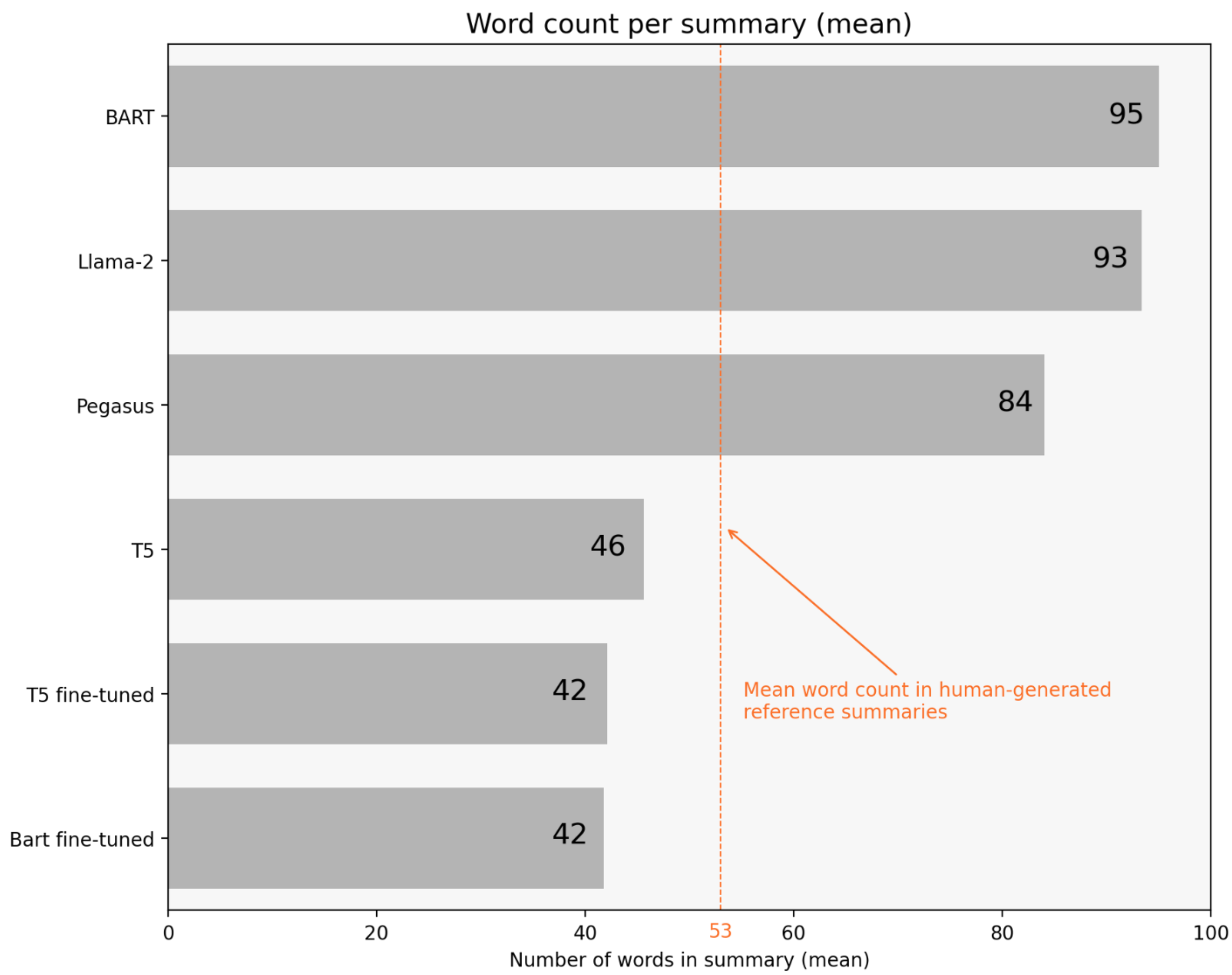

Figure 4.1 Word count per summary (mean)

BART was the most verbose model, generating summaries that were 95 words long on average. Llama-2 and Pegasus were also relatively verbose, with Llama-2 generating summaries with a mean word count of 93, and Pegasus generating summaries with a mean word count of 84. The shortest summaries were produced by the fine-tuned T5 and BART models, with a mean word count of 42. The mean word count for the human-generated reference summaries in the test dataset was 53. BART, Llama-2, and Pegasus produced summaries that were longer, on average than the human-generated summaries. T5, T5 fine-tuned, and BART fine-tuned produced summaries that were shorter, on average, than the human-generated summaries.

ROUGE Scores

For each iteration of experiments involving summary generation, the 200 model-generated summaries were compared to the corresponding reference summaries and an aggregate ROUGE score was computed for each of the following metrics: rouge1, rouge2, rougeL and rougeLsum. The aggregate ROUGE score for each ROUGE metric is the average score for the 200 records.

ROUGE scores are computed by calculating how many n-grams in the reference summary are found in the model's output summary. An n-gram is a sequence of words, where n is the length of a sequence. Rouge1 measures unigrams, or single words; rouge2 measures bi-grams, or two-word sequences; rougeL and rougeLsum measure the longest sequence of words from the reference summary that is found in the model-generated summary. RougeLsum measures the longest matching sequence of words over the whole summary, and rougeL measures the longest sequence of words for each sentence, and then returns an average for the entire summary. For rougeL and rougeLsum, the longest sequence of words does not have to be consecutive, although it must be sequential. For example, in the two sentences, "I love NY" and "I love the buildings in NY", rougeL and rougeLsum would recognize the string "I love NY" as matching.[104] ROUGE scores range from zero to one, with a lower value indicating a lack of similarity between the output summary and the reference summary, and a higher score indicating more similarity between the two.[105] The below table shows the ROUGE scores for each of the models.

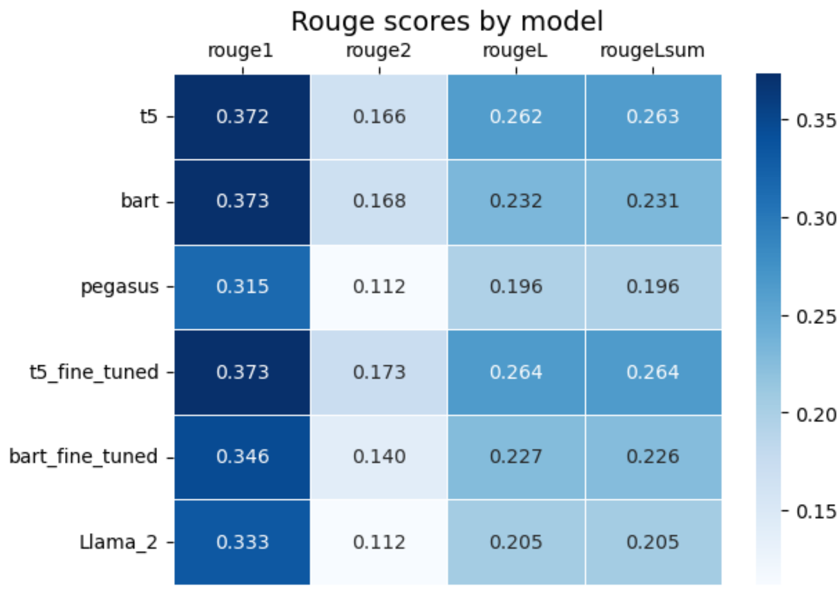

Figure 4.2 Rouge scores by model (heatmap)

[104] Alexander R. Fabbri et al., "SummEval: Re-Evaluating Summarization Evaluation," *Transactions of the Association for Computational Linguistics* 9 (2021): 393, https://doi.org/10.1162/tacl_a_00373; Chris Fregly, Antje Barth, Shelbee Eigenbrode, *Generative AI on AWS: Building Context-Aware Multimodal Reasoning Applications* (Sebastopol, CA: O'Reilly Media, Inc., December, 2023), chapter 5 "Fine-Tuning and Evaluation"; Eren Kizilirmak, "Text Summarization: How To Calculate Rouge Score," Medium, August 2, 2023, https://medium.com/@eren9677/text-summarization-387836c9e178.

[105] Janani Ravi, "Evaluation Metrics for Summaries" in "AI Text Summarization with Hugging Face," October 30, 2023, LinkedIn Learning, https://www.linkedin.com/learning/ai-text-summarization-with-hugging-face/evaluation-metrics-for-summaries.

The highest overall rouge scores of the experiment were the rouge1 scores for BART and T5 fine-tuned (.373) and T5 (.372). The lowest overall rouge scores of the experiment were the rouge2 scores for Pegasus and Llama-2, which both scored a .112. The rougeL and rougeLsum scores for each model fell between the higher rouge1 scores and the lower rouge2 scores.

The below bar chart summarizes each model's performance on the various ROUGE metrics:

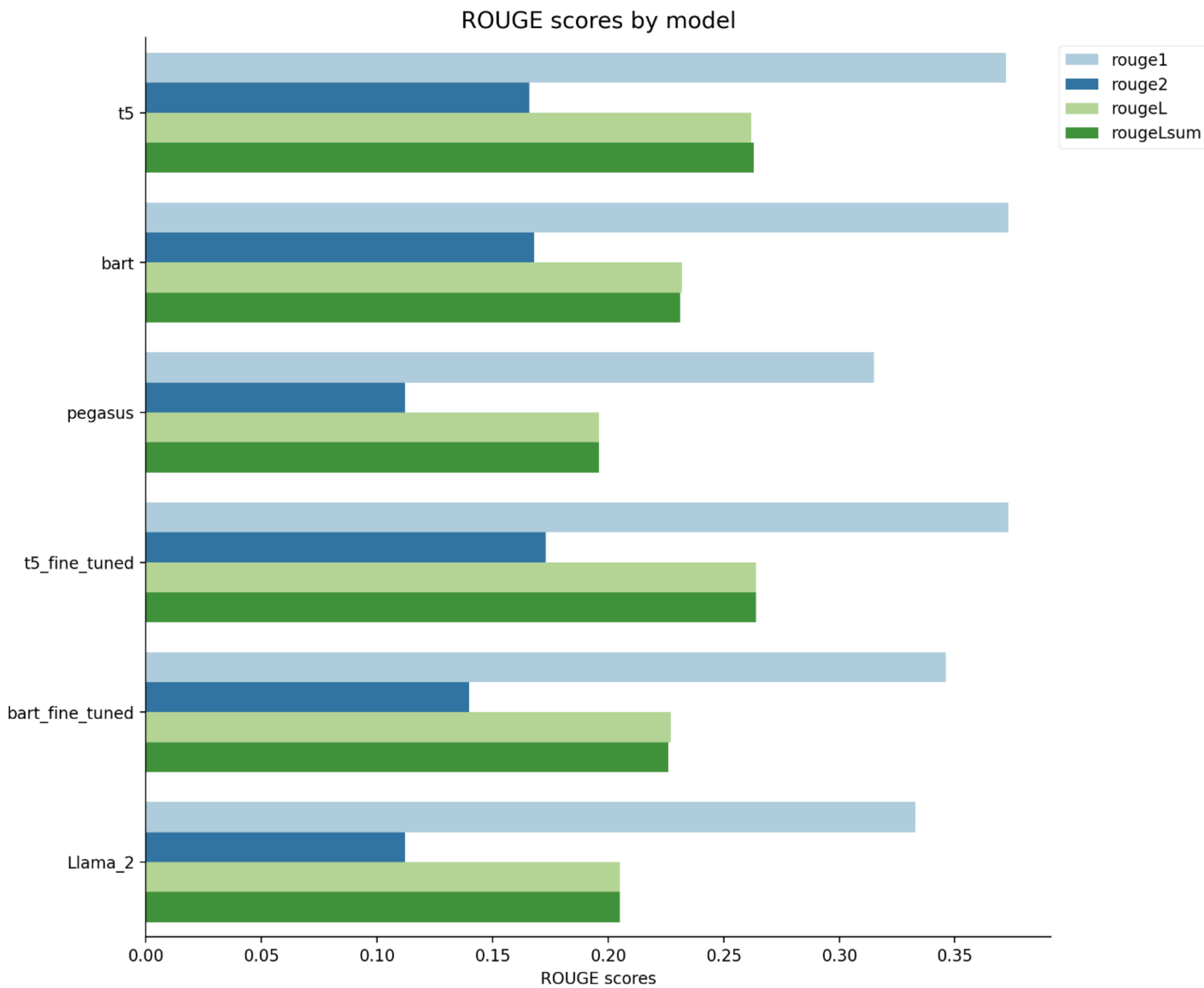

Figure 4.3 Rouge scores by model (bar plot)

Looking at each model individually, the following pattern is noted: the highest rouge scores for the model-generated summaries were achieved with the rouge1 metric; the lowest rouge scores occurred with the rouge2 metric; and the rougeL and rougeLsum metrics, which were about the same for each of the tested models, scored somewhere in between. This pattern can be observed in the below scatterplot.

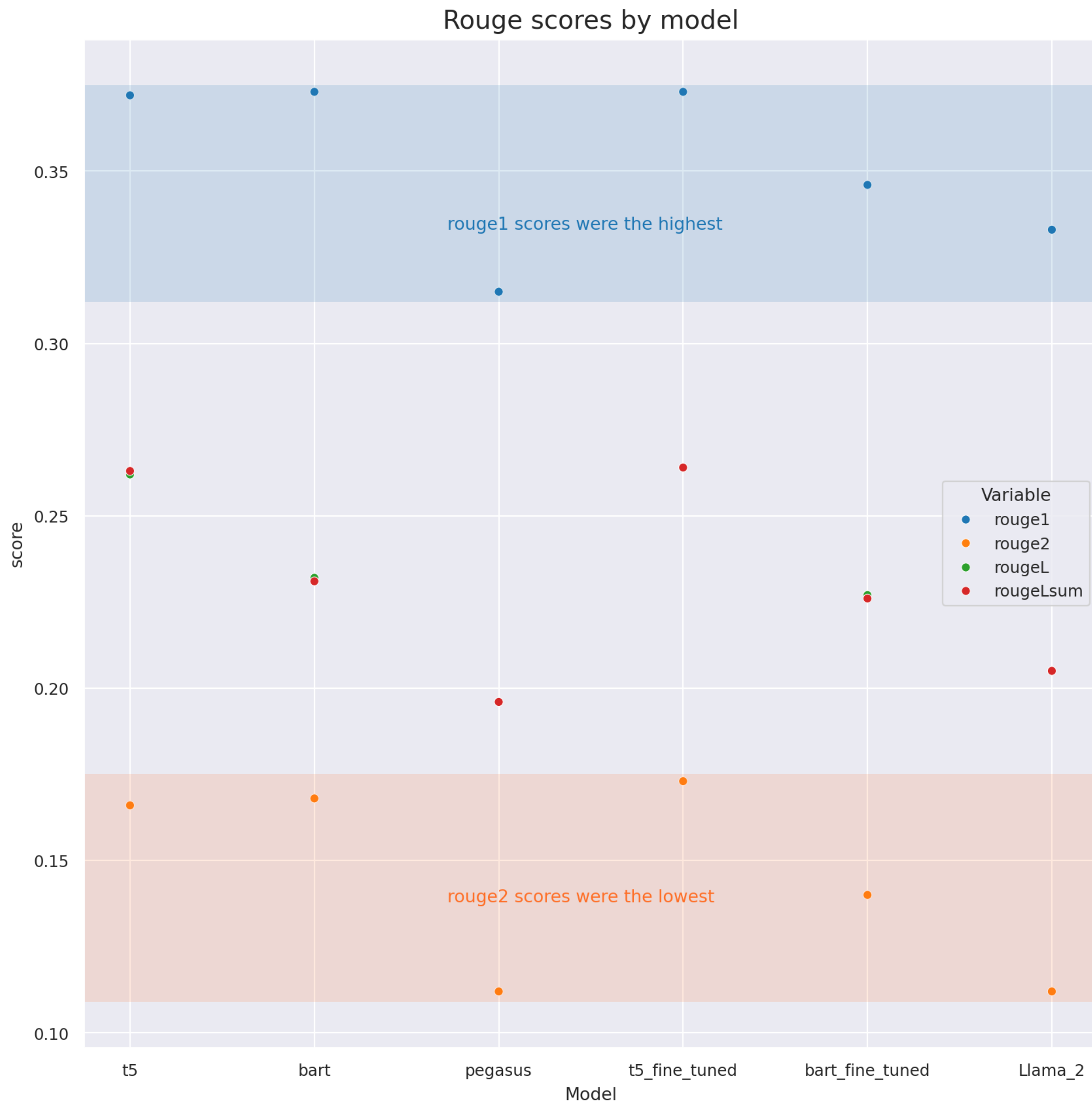

Figure 4.4 Rouge scores by model (scatterplot)

Fine-Tuning and ROUGE Scores

Experiments were conducted to understand the effect fine-tuning had on summarization quality. To measure this effect, the BART and T5 models were used to summarize the 200-article test set after which the summaries were scored using the ROUGE metrics. Both models were then fine-tuned on the 2,871 article/summary pairs in the training dataset. Next, the fine-tuned BART and T5 models were used to generate summaries for the 200-article test set and these summaries were again scored using the ROUGE metrics. The fine-tuned T5 model outscored the T5 model in every ROUGE metric, however the magnitude of the increase was relatively small. For example, the T5 fine-tuned model scored a .373 on the rouge1 metric, compared to the off-the-shelf T5 model's score of .372. The BART fine-tuned model scored lower than the BART model in every ROUGE metric, but the magnitude of the decrease was not significant. For example, the BART fine-tuned model scored a .227 on the rougeL score whereas the BART model scored a .232 on rougeL. The below table and slope graphs illustrate the impacts of fine-tuning on ROUGE scores.

Model	rouge1	rouge2	rougeL	rougeLsum
T5	.372	.166	.262	.263
T5 fine-tuned	.373	.173	.264	.264
BART	.373	.168	.232	.231
BART fine-tuned	.346	.140	.227	.226

Table 4.1 Impact of fine-tuning on ROUGE scores

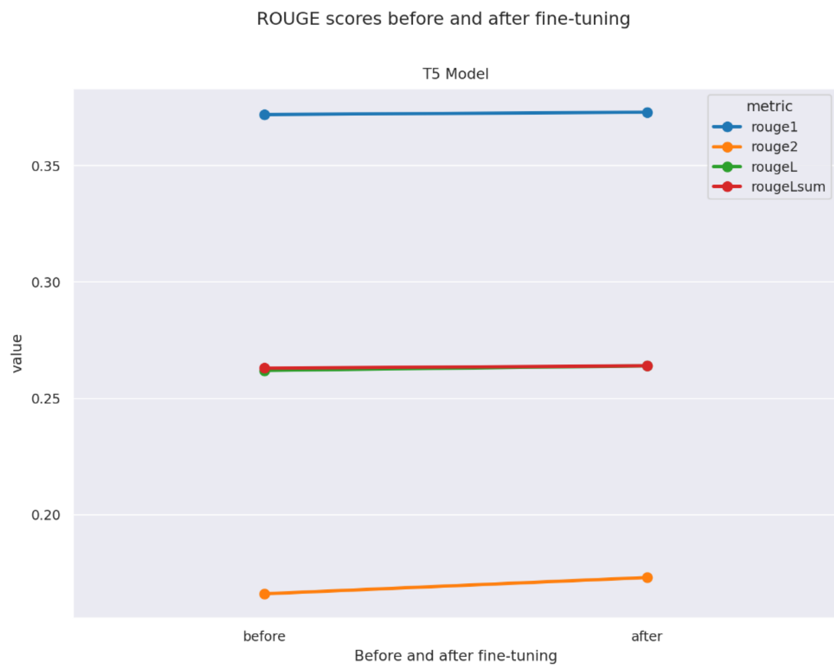

Figure 4.5 ROUGE scores before and after fine-tuning T5

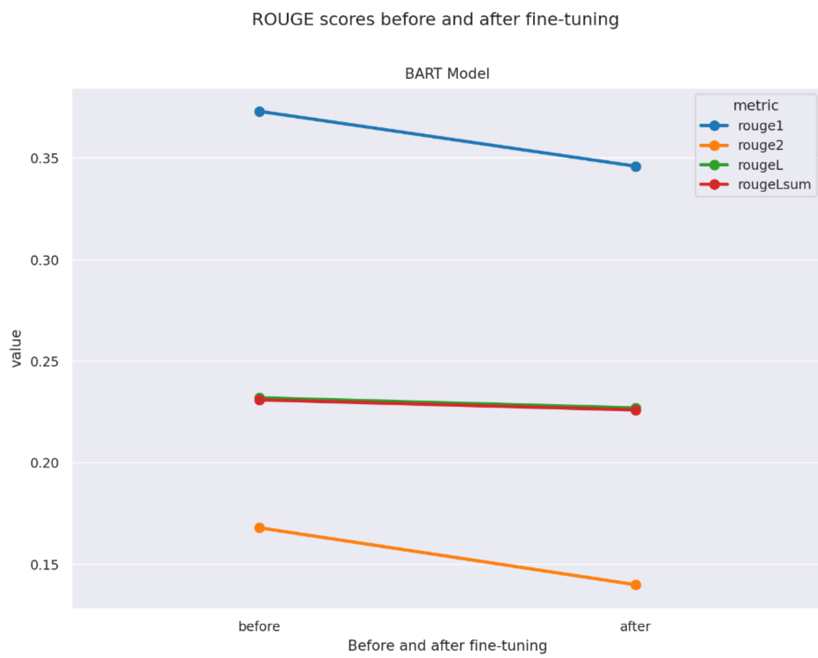

Figure 4.6 ROUGE scores before and after fine-tuning BART

Hyperparameter Testing

T5 was subjected to an array of experiments to test how changes to three hyperparameters (learning rate, batch size, and training epochs) during the training phase affected ROUGE scores. ROUGE scores were calculated after each hyperparameter change. The ROUGE scores were then compared to the off-the-shelf T5 model which had not been fine-tuned. Learning rate determines the step size that is used during model training and optimization and dictates how the model weights are adjusted to minimize error and improve performance. The learning rate settings were tested at .001, .01, and .1. When learning rates were set at .001 and .01, the results were very similar to the baseline model. However, when learning rate was set to .1, the ROUGE scores decreased significantly: rouge1 was .067, rouge2 was 0, and rougeL was .06. This demonstrates the impact that learning rate can have on model performance and the importance of experimenting with changes to learning rate during model training. The below plot summarizes the findings of the learning rate experiments.

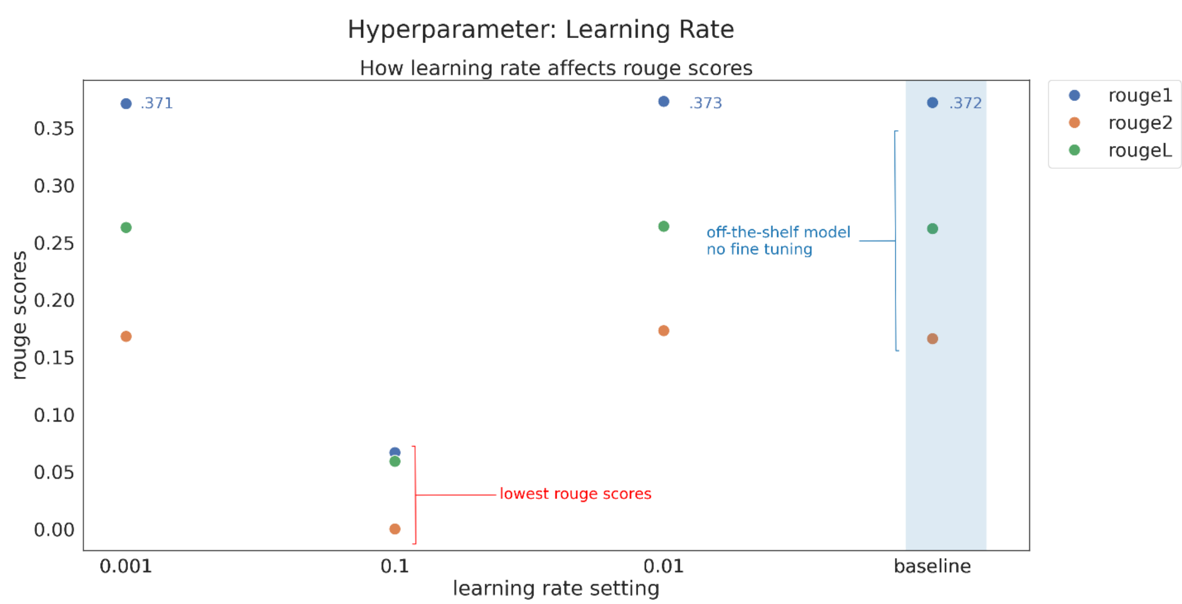

Figure 4.7 Hyperparameter Testing: Learning Rate and ROUGE Score

The batch size settings were tested at 8, 16, and 32. Batch size refers to how the entire dataset is divided into smaller portions called batches when machine learning models are trained. The model weights are adjusted after each batch using an optimization function, such as gradient descent. The batch size hyperparameter determines how many examples are included in each batch.[106] Changing the batch size did not significantly change ROUGE scores – the changes only impacted ROUGE scores by a few hundredths. The below plot summarizes the findings of the batch size experiments.

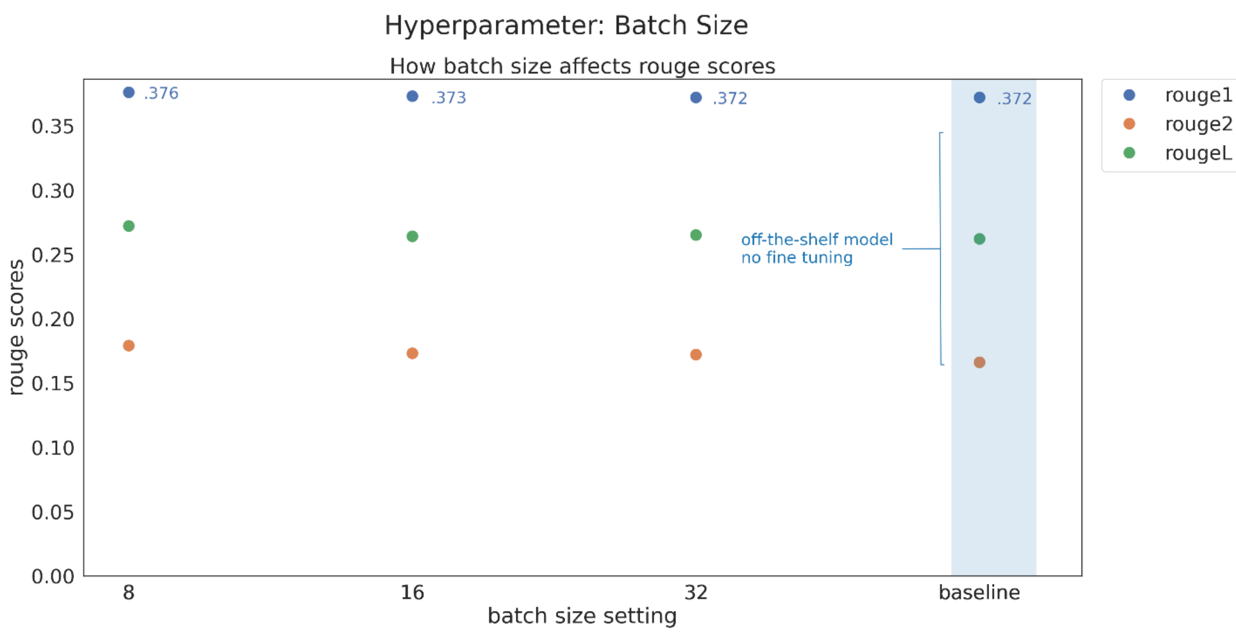

Figure 4.8 Hyperparameter Testing: Batch Size and ROUGE Score

The training epochs settings were tested at 2, 4, and 8. Training epoch is one complete cycle of ML model training. For example, after all batches of training data are passed to a

[106] Delip Rao and Brian McMahan, *Natural Language Processing with PyTorch* (Sebastopol, CA: O'Reilly Media, Inc., 2019), chapter 3.

model, the training dataset is complete and one training epoch has occurred. The number of training epochs refers to how many iterations of training a model undergoes.[107] Altering the number of training epochs did not significantly change ROUGE scores – the changes only impacted ROUGE scores by a few hundredths. The below plot summarizes the findings of the training epoch experiments.

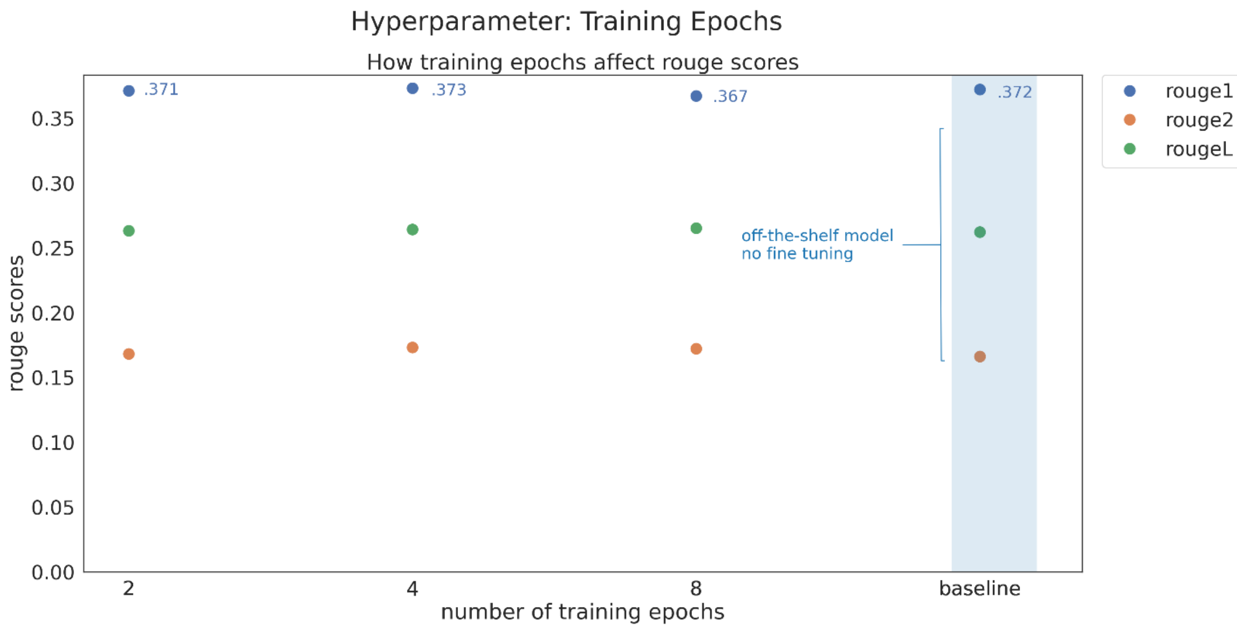

Figure 4.9 Hyperparameter Testing: Training Epochs and ROUGE Score

Llama3 Evaluation of Summary Quality

Meta's LLama-3 70 billion parameter LLM was used as to evaluate summary quality using a qualitative approach. The purpose of this experiment was to augment the quantitative ROUGE scores with a qualitative evaluation that could, to an extent, emulate human evaluation. The model was provided with a news article and a series of model-generated summaries and was

[107] Rao and McMahan, *Natural Language Processing with PyTorch*, chapter 3.

instructed to select the best summary. The model was instructed as follows: "You are an expert evaluator for judging summaries. You will be given an article and [a set of] different summaries. Your task is to choose which summary is the best." The Llama-3 model was used to compare model-generated summaries across the following dimensions for each record in the 200-article test set.

BART vs. T5 vs. Pegasus

Llama3 was instructed to choose the best summary from the off-the-shelf models. These models were not fine-tuned. The purpose of this experiment was to identify which model Llama3 judged to be the best in an off-the-shelf configuration with no fine-tuning. To control for the possibility that Llama3's selection of the best summary might be influenced by the order in which the three summaries were presented to it, the experiment was run six different times to account for each possible permutation of orders (BART/Pegasus/T5; BART/T5/Pegasus; Pegasus/BART/T5; Pegasus/T5/BART; etc.). In essence, Llama3 voted six times on which of the three summaries it thought was best (one vote for each possible order). If Llama3 voted for the summary of a particular model more than three times, a simple majority was achieved, and that model was identified as the "winner." If the voting after six iterations was tied (either 2-2-2 or 3-3), the results were indecisive. For the 200 records in the test dataset, Llama3 voted as follows:

LLM selection of best summary	Count (out of 200)	Percentage (%)
BART	116	58%
Pegasus	41	20.5%
T5	18	9%
Voting tied (no best model, results indecisive)	25	12.5%

Table 4.2 Llama3 LLM as Evaluator: Bart vs. T5 vs. Pegasus (off-the shelf, no fine-tuning)

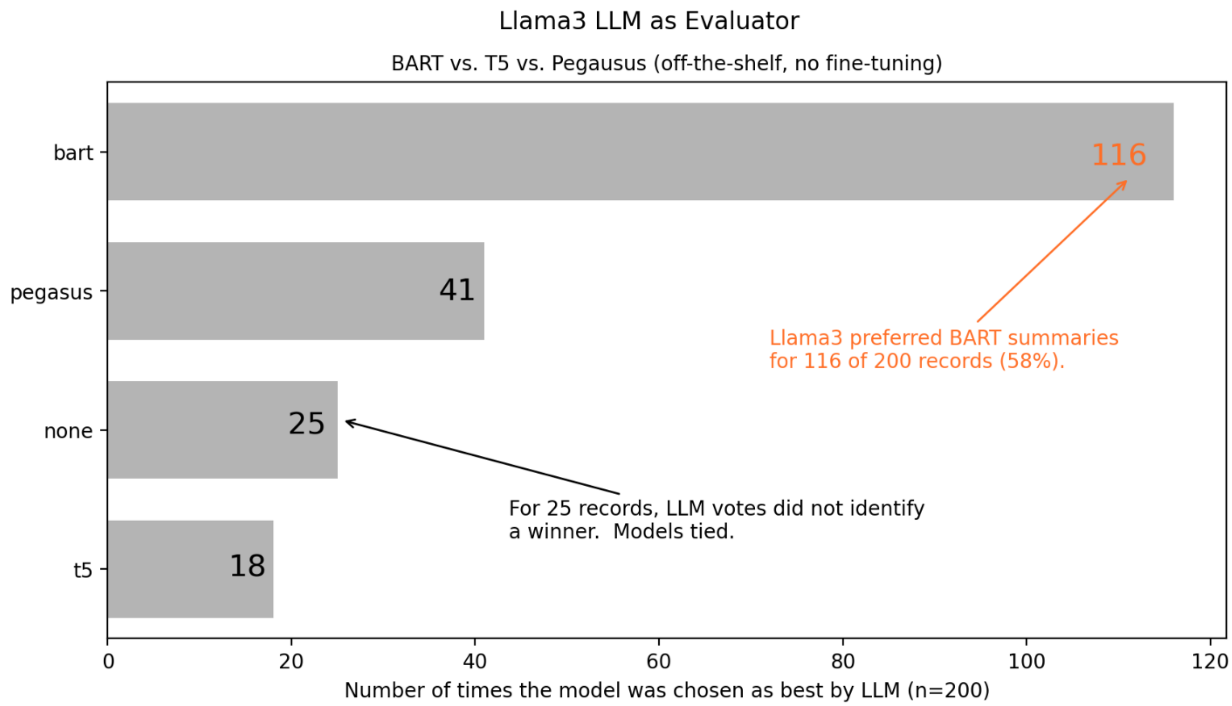

Figure 4.10 Llama3 LLM as Evaluator: Bart vs. T5 vs. Pegasus (off-the shelf, no fine-tuning)

BART fine-tuned vs. T5 fine-tuned

Llama3 was instructed to choose the better summary between those generated by the T5 fine-tuned model and the BART fine-tuned model. The purpose of this experiment was to identify which of the fine-tuned models Llama3 judged to be the best. To control for the possibility that Llama3's selection of the best summary might be influenced by the order in which the two summaries were presented to it, the experiment was run six separate times for two reasons. The first reason is that doing so allowed the LLM to evaluate the summaries in each of the two possible order permutations (BART fine-tuned/T5 fine-tuned; T5 fine-tuned/BART fine-tuned). The second reason for running the experiment six times was to allow the test conditions of this experiment to match the test conditions of the first LLM experiment in which six iterations were necessary to account for each possible permutation of the three off-the-shelf models (BART vs. Pegasus vs. T5). In essence, Llama3 voted six times on which of the two summaries it thought was best (one vote for each iteration of the experiment). If Llama3 voted for the summary of a particular model more than three times, a simple majority was achieved, and that model was identified as the "winner." If the voting after six iterations was tied (three votes for the BART fine-tuned model and three votes for the T5 fine-tuned model), the results were indecisive. For the 200 records in the test dataset, Llama3 voted as follows:

LLM selection of best summary	**Count (out of 200)**	**Percentage (%)**
BART fine-tuned	97	48.5%
T5 fine-tuned	42	21%
Voting tied (no best model, results indecisive)	61	30.5%

Table 4.3 Llama3 LLM as Evaluator: BART fine-tuned vs T5 fine-tuned

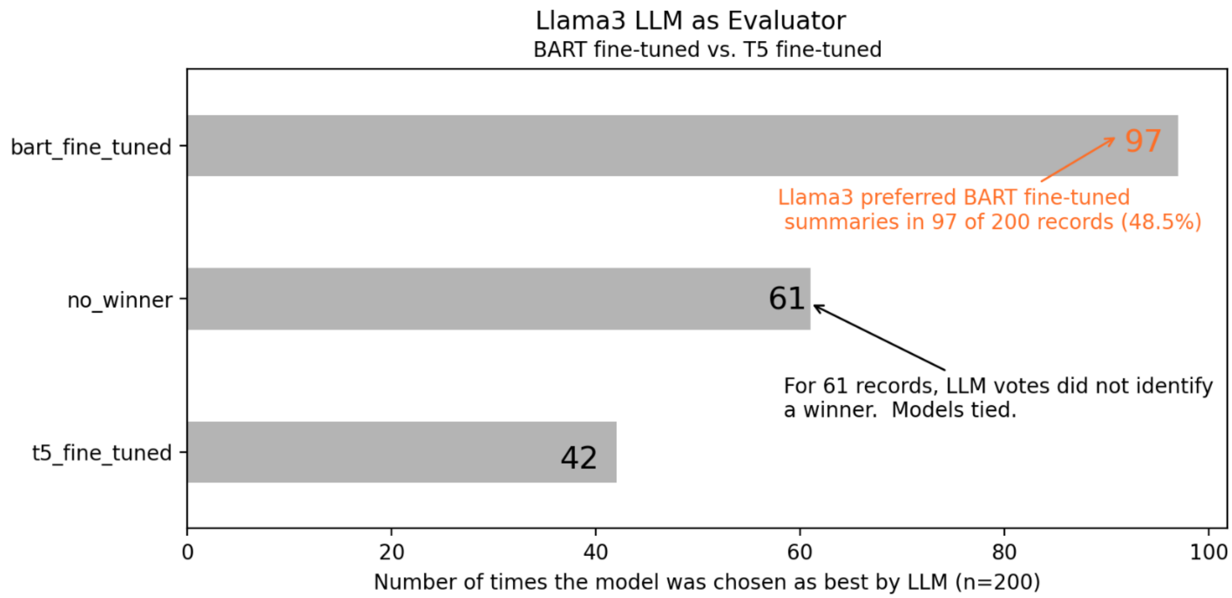

Figure 4.11 Llama3 LLM as Evaluator: BART fine-tuned vs. T5 fine-tuned

BART vs. BART fine-tuned

Llama3 was instructed to choose the better summary between those generated by the BART model and the BART fine-tuned model. The purpose of this experiment was to identify whether Llama3 preferred the summaries produced by the BART model or the BART fine-tuned model. To control for the possibility that Llama3's selection of the best summary might be influenced by the order in which the two summaries were presented to it, the experiment was run six separate times for two reasons. The first reason is that doing so allowed the LLM to evaluate the summaries in each of the two possible order permutations (BART/BART fine-tuned; BART fine-tuned/BART). The second reason for running the experiment six times was to allow the test conditions of this experiment to match the test conditions of the first LLM experiment in which six iterations were necessary to account for each possible permutation of the three off-the-shelf models (BART vs. Pegasus vs. T5). In essence, Llama3 voted six times on which of the two summaries it thought was best (one vote for each iteration of the experiment). If Llama3 voted for the summary of a particular model more than three times, a simple majority was achieved, and that model was identified as the "winner." If the voting after six iterations was tied (three votes for the BART model and three votes for the BART fine-tuned model), the results were indecisive. For the 200 records in the test dataset, Llama3 voted as follows:

LLM selection of best summary	Count (out of 200)	Percentage (%)
BART	95	47.5%
BART fine-tuned	59	29.5%
Voting tied (no best model, results indecisive)	46	23%

Table 4.4 Llama3 LLM as Evaluator: BART vs. BART fine-tuned

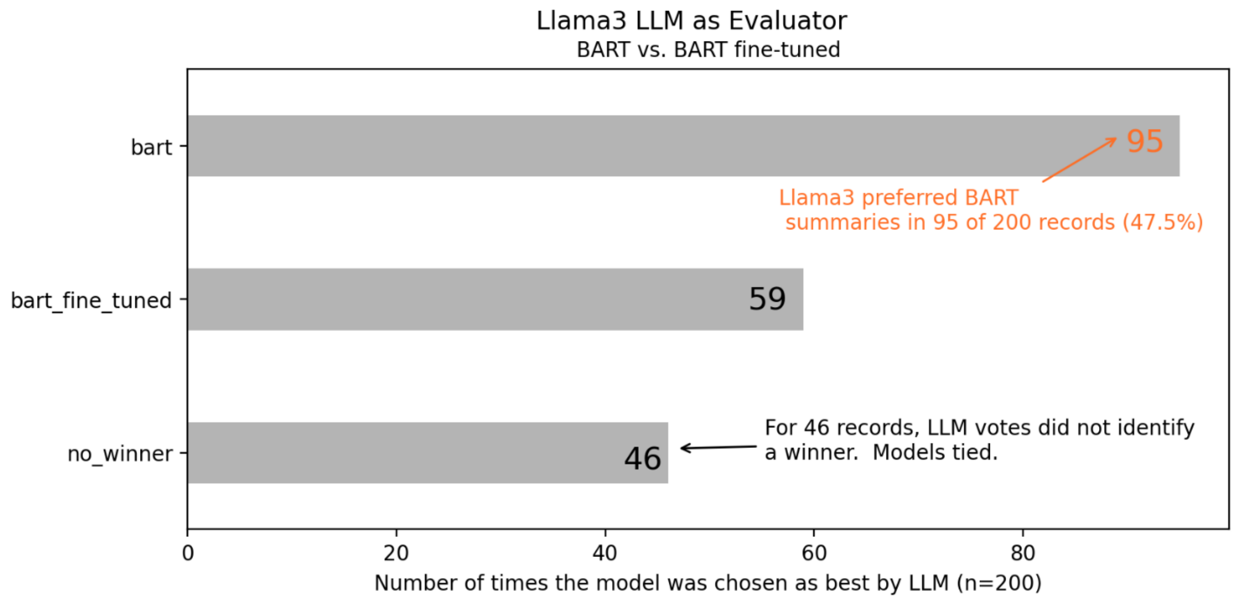

Figure 4.12 Llama3 LLM as Evaluator: BART vs. BART fine-tuned

T5 vs. T5 fine-tuned

Llama3 was instructed to choose the better summary between those generated by the T5 model and the T5 fine-tuned model. The purpose of this experiment was to identify whether Llama3 preferred the summaries produced by the T5 model or the T5 fine-tuned model. To control for the possibility that Llama3's selection of the best summary might be influenced by the order in which the two summaries were presented to it, the experiment was run six separate times for two reasons. The first reason is that doing so enables the LLM to evaluate the summaries in each of the two possible order permutations (T5/T5 fine-tuned; T5 fine-tuned/T5). The second reason for running the experiment six times was to allow the test conditions of this experiment to match the test conditions of the first LLM experiment in which six iterations were necessary to account for each possible permutation of the three off-the-shelf models (BART vs. Pegasus vs. T5). In essence, Llama3 voted six times on which of the two summaries it thought was best (one vote for each iteration of the experiment). If Llama3 voted for the summary of a particular model more than three times, a simple majority was achieved, and that model was identified as the "winner." If the voting after six iterations was tied (three votes for the T5 model and three votes for the T5 fine-tuned model), the results were indecisive. For the 200 records in the test dataset, Llama3 voted as follows:

LLM selection of best summary	Count (out of 200)	Percentage (%)
T5 fine-tuned	95	47.5%
T5	63	31.5%
Voting tied (no best model, results indecisive)	42	21%

Table 4.5 Llama3 LLM as Evaluator: T5 vs. T5 fine-tuned

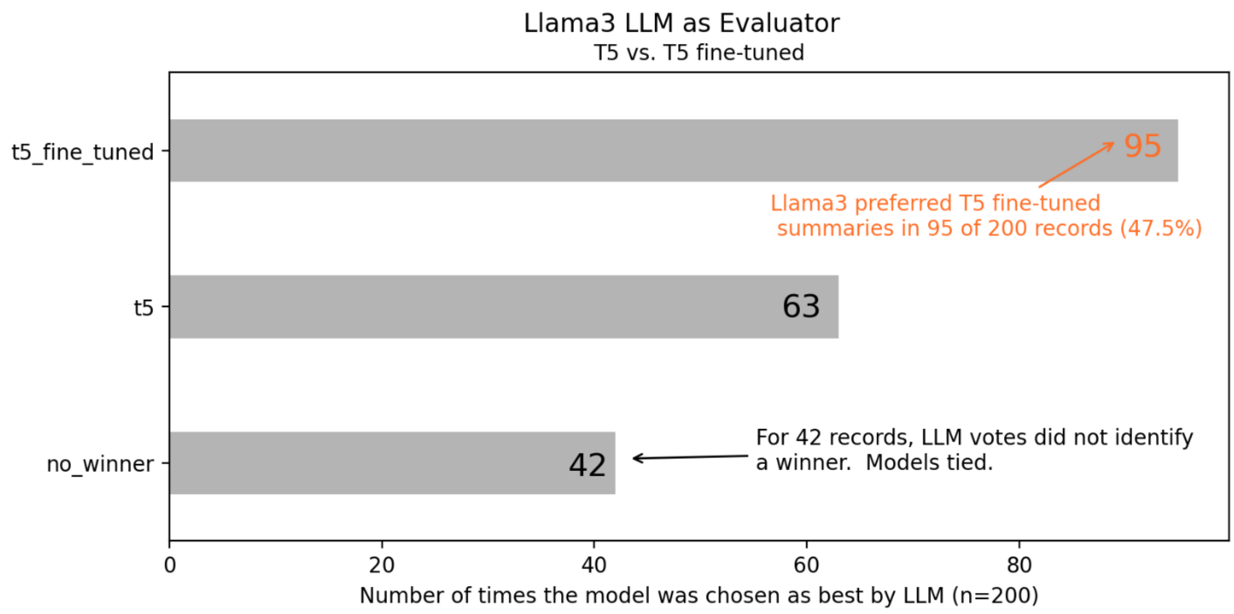

Figure 4.13 Llama3 LLM as Evaluator: T5 vs. T5 fine-tuned

Comparison of all models: BART, BART fine-tuned, T5, T5 fine-tuned, Pegasus, Llama2

Llama3 was instructed to select the best summary from a set of summaries generated by all six of the models in the experiment: BART, BART fine-tuned, T5, T5 fine-tuned, Pegasus, Llama2. The purpose of this experiment was to identify which summary Llama3 preferred from all model variants including off-the shelf models with no fine-tuning (BART, T5, and Pegasus), fine-tuned models (BART fine-tuned and T5 fine-tuned), and the LLM (Llama2). To control for the possibility that Llama3's selection of the best summary might be influenced by the order in which the six summaries were presented to it, the experiment was run six separate times for two reasons. The first reason is that doing so enabled Llama3 to evaluate the summaries in six possible order permutations. The second reason for running the experiment six times was to allow the test conditions of this experiment to match the test conditions of the first LLM experiment in which six iterations were necessary to account for each possible permutation of the three off-the-shelf models (BART vs. Pegasus vs. T5). In essence, Llama3 voted six times on which of the six summaries it thought was best (one vote for each iteration of the experiment). If Llama3 voted for the summary of a particular model more than three times, a simple majority was achieved, and that model was identified as the "winner." If the voting after six iterations was tied (for example, one vote for each of the six summaries, or three votes for one summary and three votes for another), the results were indecisive. For the 200 records in the test dataset, Llama3 voted as follows:

LLM selection of best summary	Count (out of 200)	Percentage (%)
Llama2	196	98%
Pegasus	1	0.5%
BART	1	0.5%
BART fine-tuned	0	0%
T5	0	0%
T5 fine-tuned	0	0%
Voting tied (no best model, results indecisive)	2	1%

Table 4.6 Llama3 LLM as Evaluator: Comparison of all models

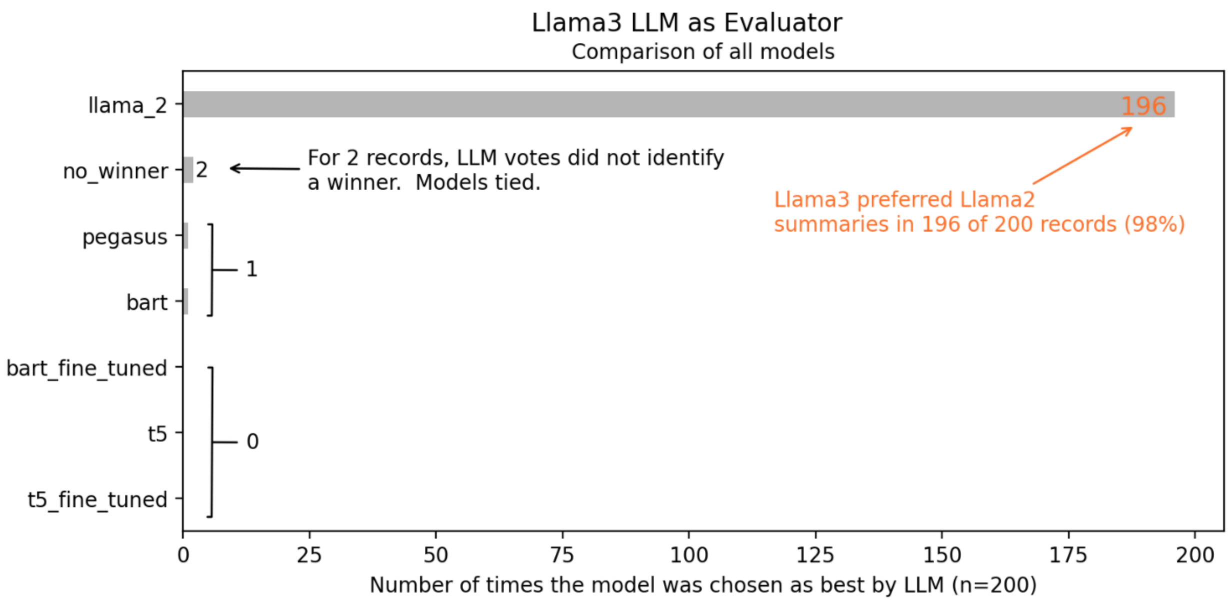

Figure 4.14 Llama3 LLM as Evaluator: Comparison of all models

The below table consolidates the results of all five experiments in which Llama3 served as an evaluator:

Experiment	**Top model chosen by Llama3**	**Percentage of vote**	**Second-best model chosen by Llama3**	**Percentage of vote**
All models (Llama2, BART, BART fine-tuned, T5, T5 fine-tuned, Pegasus)	Llama2	98%	Pegasus and Bart	0.5%
Off-the-shelf models (BART, T5, Pegasus)	BART	58%	Pegasus	20.5%
Fine-tuned (BART fine-tuned, T5 fine-tuned)	BART fine-tuned	48.5%	T5 fine-tuned	21%
BART only (BART, BART fine-tuned)	BART	47.5%	BART fine-tuned	29.5%
T5 only (T5, T5 fine-tuned)	T5 fine-tuned	47.5%	T5	31.5%

Table 4.7 Llama3 LLM as Evaluator: Consolidated results

When all models were compared, Llama3 had a strong preference for the Llama2 summaries and selected them as the best 98% of the time. When the off-the-shelf models were compared, Llama3 selected the BART summaries as the best 58% of the times, beating out Pegasus and T5 by a significant margin. When comparing summaries generated by the T5 fine-tuned model and the BART fine-tuned model, Llama3 preferred the BART fine-tuned summaries 48.5% of the time, compared to the T5 fine-tuned summaries which Llama3 only selected 21% of the time. Llama3 was used in two experiments to evaluate the extent to which fine-tuning a model improved the summaries it generated. This was the BART only experiment in which the BART model was compared to the BART fine-tuned model; and the T5 only experiment in which the T5 model was compared to the T5 fine-tuned model. For BART, Llama3 preferred the off-the shelf model over the fine-tuned model (47.5% to 29.5%) . For T5, Llama3 preferred the fine-tuned model over the off-the-shelf model (47.5% to 31.5%).

CHAPTER 5

Discussion of Results and Conclusion

Summary of Conclusions

The research experiments support the following conclusions, which are further discussed below:

- **ROUGE metric has limited utility for evaluating abstractive summaries** – The ROUGE metric measures word and phrase overlap between reference summaries and model-generated summaries. The semantic similarity between two bodies of text cannot be accurately measured by word and phrase overlap alone because today's models use generative AI and similar ideas can be expressed with different words. Therefore, text summarization researchers should carefully evaluate whether ROUGE is a useful metric for a study's objectives. There is consensus among contemporary researchers that ROUGE has limited usefulness, and this study's findings support that position.[108]
- **ROUGE could not be optimized for based on the experiments –** Fine-tuning BART and T5 did not markedly improve ROUGE scores. This may be because the training dataset of 2,871 records was too small to change the model. Changes to hyperparameter settings during fine-tuning did not markedly improve ROUGE performance either.
- **Llama3 is a useful evaluator of summaries –** The Llama3 LLM was used to evaluate the quality of summarizations. Llama3 offers greater flexibility compared to ROUGE

[108] Fabian Retkowski, "The Current State of Summarization," in Beyond Quantity: Research with Subsymbolic AI (forthcoming, November 2023; published online August 1, 2023): 4, http://arxiv.org/abs/2305.04853; Tong Chen et al., "Enhancing Abstractive Summarization with Extracted Knowledge Graphs and Multi-Source Transformers," *Applied Sciences* 13, no. 13 (June 2023): 10, https://doi.org/10.3390/app13137753.

because it enables simultaneous comparison of multiple bodies of text, whereas ROUGE is limited to comparing only two texts at a time. Despite Llama3's flexibility, there is some uncertainty about whether Llama3's assessments of summarization quality are reliable. Nonetheless, it is an improvement over ROUGE. The use of LLMs like Llama3 as an evaluator of summarization quality is an area of active research.[109]

- **Llama2 LLM performed better than other models** – Based on the author's qualitative evaluation of model-generated summaries for six records, the Llama2 summaries seem to read better and are more complete than the summaries produced by the BART, T5, and Pegasus variants. This finding accords with contemporary research showing that LLMs outperform pre-trained language models like BART and T5 on the summarization task.[110] However, the sample size of the qualitative evaluation for this study was quite small, and human evaluation is subject to various discrepancies and inconsistencies.[111] When deploying LLMs for summarization tasks, IC organizations should be aware of the risks of hallucinations and take steps to offset those risks such as keeping a human-in-the-loop.[112] IC organizations should also understand that technical challenges exist when

---

[109] Tianyi Zhang et al., "Benchmarking Large Language Models for News Summarization," *Transactions of the Association for Computational Linguistics* 12 (2024): 39-57, https://doi.org/10.1162/tacl_a_00632; Pat Verga et al., "Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models," (Ithaca: Cornell University Library, arXiv.org, April 29, 2024): 1, https://doi.org/10.48550/arXiv.2404.18796.

[110] Xiao Pu, Mingqi Gao, and Xiaojun Wan (Wangxuan Institute of Computer Technology, Peking University), "Summarization is (Almost) Dead," working paper (Ithaca: Cornell University Library, arXiv.org, September, 2023): 1-2, 4. https://doi.org/10.48550/arXiv.2309.09558.

[111] Inderjeet Mani, *Automatic Summarization* (Philadelphia: John Benjamins Publishing Company, 2001), 225.

[112] Tong Chen et al., "Enhancing Abstractive Summarization with Extracted Knowledge Graphs and Multi-Source Transformers," *Applied Sciences* 13, no. 13: 7753 (June 2023): 1, https://doi.org/10.3390/app13137753.

bringing in an LLM to a secure, air-gapped environment. While significant, these challenges are not insurmountable and have already been solved within various parts of the IC.

- **Measuring summarization quality is inherently difficult** – Summaries are context-dependent and serve different use cases in different settings. Although there is currently no widely accepted automated metric to measure summarization quality, there are several approaches and frameworks that may be useful. One promising approach is to identify the key points that must be included in a summary for a particular use case, then testing whether the model-generated summary includes the key points.
- **Organizational culture can impact the success of AI/ML initiatives in the IC –** Organizational culture, the resources that IC leaders allocate to AI/ML initiatives, and data availability are important factors for IC organizations that seek to deploy production-capable AI/ML pipelines at scale. This holds true for summarization pipelines.

ROUGE Metric is of Limited Use for Evaluating Abstractive Summaries

A central question of this thesis is: which model is best for summarization? Organizations within the IC and national security enterprise that aspire to implement language models for summarization tasks will rightfully seek reliable answers to this question. Similarly, organizations that have already instantiated summarization pipelines may wonder how well the models are performing, or whether the models in use are the most performant ones for the task. Although there are several ways to approach these concerns, the research in this thesis supports the position that summarization pipelines must be looked at from two interrelated perspectives: the quality of the summaries being generated, and the characteristics of the systems and technology that are used to evaluate the summaries. Without a reliable, automated evaluation system, organizations will default to human testing and judgement which is time-consuming, expensive, subjective, and unscalable.[113] Hence, the tendency within the text summarization research community has been to use automated metrics, at least partially, to assess model performance.[114]

Because of its purported capability to provide a degree of quantitative measurement of summarization quality, several of the experiments in this thesis attempted to optimize for the ROUGE metric. The BART, T5, and Pegasus model were tested in an off-the-shelf configuration, and the highest score achieved was a .373 rouge1 score (ROUGE metric ranges from 0 to 1, with 1 implying the highest similarity). Llama 2, an LLM, was also tested on the summarization task, and the highest score it received was a rouge1 score of .333. The T5 and BART models were also fine-tuned to test the effect fine-tuning would have on ROUGE scores. After fine-tuning T5, the ROUGE scores for the fine-tuned T5 model increased by a few thousandths for each ROUGE category. Since the increase of the ROUGE scores for the T5 fine-tuned model as compared to the T5 model was so small, a conclusion of this research is that fine-tuning the T5 model did not provide a significant improvement to ROUGE scores.
After fine-tuning BART, the ROUGE scores for the fine-tuned BART model decreased for every ROUGE category by an average of .039. One possible explanation of why the ROUGE scores

[113] Chin-Yew Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out* (Barcelona, Spain: Association for Computational Linguistics, 2004): 1, https://aclanthology.org/W04-1; Inderjeet Mani, *Automatic Summarization* (Philadelphia: John Benjamins Publishing Company, 2001), 225.

[114] Vivek Srivasta, Savita Bhat, and Niranjan Pedanekar, "Hiding in Plain Sight: Insights into Abstractive Text Summarization," *The Fourth Workshop on Insights from Negative Results in NLP, Association of Computational Linguistics* (2023): 68-70, http://dx.doi.org/10.18653/v1/2023.insights-1.8

for the fine-tuned BART model were lower than the off-the-shelf BART model is that the mean word count for the BART fine-tuned summaries decreased by nearly 56%, from 95 to 42, as compared to the BART summaries. Since ROUGE scores are computed based on word and phrase overlap between the reference summary and the generated summary, a model-generated summary with fewer words is less likely to share word and phrase overlap with the reference summary compared to a longer model-generated summary. Therefore, the lower ROUGE scores for the BART fine-tuned model may be explained by the fact that the fine-tuned summaries were significantly shorter than the summaries generated by the off-the-shelf model. On the other hand, the T5 fine-tuned summaries were only about 7% shorter than the T5 off-the-shelf model summaries, which may partially explain why the ROUGE scores for the T5 fine-tuned summaries were in fact slightly better than those produced by the T5 off-the-shelf model. While the word count of the T5 and BART summaries can be controlled through hyperparameter adjustments, those changes were not part of the experiments conducted for this research.

A battery of hyperparameter experimentation was conducted on T5 to determine the extent to which changes to learning rate, batch size, and training epochs impacted ROUGE scores. Given the test conditions for this research, changes to the batch size and number of training epochs did not impact ROUGE scores by more than a few hundredths. That was mostly true for learning rate as well, except for one experiment when the ROUGE scores dropped significantly after learning rate was set to .1 (rouge1 went from .372 to .067, and rouge2 went from .166 to 0). These results support a conclusion that it was not possible to optimize for the ROUGE metric given the models, dataset, and conditions of the experiments. While it might be possible that the training dataset was too small to change model performance during fine-tuning and hyperparameter testing, this thesis did not examine the effects a larger training dataset would have on the ROUGE scores.

To provide a level of ground truth to the ROUGE scores, we can qualitatively examine a few of the high and low-scoring summaries. Although the T5 rouge1 aggregate score, which is an average, was .372, several T5 summaries scored significantly higher on an individual basis. For example, the following T5-generated summary, which relates to the results of a Chelsea soccer match, scored .623 on the rouge1 metric:

> *chelsea beat maribor 7-0 in the uefa youth league on wednesday afternoon . dominic solanke and charlie colkett scored two goals apiece . charly musonda, alex kiwomya and tammy abraham also scored for the blues .*

The human-generated reference summary, against which the T5 summary was scored, reads as follows:

> *chelsea beat maribor 7-0 to extend lead in uefa youth league group .dominic solanke and charlie colkett scored twice each in slovenia .charly musonda, alex kiwomya and tammy abraham also scored .result means chelsea preserve their 100 per cent group stage record .*

The words in the model-generated summary appear frequently in the reference summary, especially the names. This condition resulted in the high rouge1 score. In this instance, the model-generated summary approximates the reference summary rather well. On the other hand, the following T5-generated summary, which relates to alleged war crimes in Syria, scored .154 on the rouge1 metric:

u.n., qatar and saudi arabia have compiled a report into syria . a spokesman said the images were 'extremely disturbing' and 'horrible to look at'

The human-generated summary, against which the T5 summary was scored, reads as follows:

warning: graphic content .witness said he was tasked to record deaths in custody from 2011 to 2013 .there are 55,000 photos which lawyers say are evidence of extreme torture .report made by ex-war crimes prosecutors who deem witness 'credible'one lawyer said abuses are likely to be even more extensive .foreign secretary said the images are 'compelling and horrific'dossier commissioned by british lawyers for qatar which supports rebels .assad regimes questioned the authenticity of the photographs .

The words in the T5-generated summary do not align with the reference summary, resulting in the low ROUGE score. These examples show that the ROUGE scores reliably predict the level of word overlap between model-generated summary and reference summary. But there is a more fundamental question that ROUGE does not measure - how *well* does the model-generated summary capture the key information in the news article? This is one of several shortcomings of the ROUGE metric.

Although ROUGE has been a standard metric for measuring summarization, this thesis supports the consensus held by contemporary text summarization researchers that the ROUGE metric is faulty.[115] In fact, this thesis concludes that the ROUGE metric may be nonviable for evaluating model performance. There are several reasons for this conclusion. One of the main

[115] Vivek Srivasta, Savita Bhat, and Niranjan Pedanekar, "Hiding in Plain Sight: Insights into Abstractive Text Summarization," *The Fourth Workshop on Insights from Negative Results in NLP, Association of Computational Linguistics* (2023): 68-70, http://dx.doi.org/10.18653/v1/2023.insights-1.8; Tong Chen et al., "Enhancing Abstractive Summarization with Extracted Knowledge Graphs and Multi-Source Transformers," *Applied Sciences* 13, no. 13 (June 2023): 10, https://doi.org/10.3390/app13137753.; Retkowski, "The Current State of Summarization," 4.

reasons is that today's top NLP models are generative, meaning they create new language and ideas from inputs. In the case of summarization, this means that a model generates a summary not by taking sentences out of a body of text and rearranging them, but by understanding semantic meaning and generating new words and sentences to summarize the input text. This is the heart of generative AI, and this trend will accelerate as massive, powerful models continue to come online.[116] In the context of text summarization, this implies that users will generally expect abstractive summaries, which the models will predominantly provide by default. Problematically, ROUGE was designed for an outdated family of language models that created extractive summaries by taking words, word groups, and sentences from an input and rearranging them as an output. ROUGE measures summarization performance by looking for word-gram overlap in the model-generated and reference text. ROUGE scores increase as the metric detects words and phrases in the generated text that also appear in the reference text. Conversely, ROUGE scores decrease as the metric detects less overlap with the reference input. Therefore, the fundamental, yet flawed idea behind the ROUGE metric is that semantic similarity between two bodies of text increases as the number of word-grams increase. The inaccuracy of this premise can easily be demonstrated with two basic examples which show how ROUGE scores tell us nearly nothing about the semantic similarity between two sentences:

---

[116] Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," *Applied Sciences 13*, no. 13: 7620 (June 2023): 1-20, https://doi.org/10.3390/app13137620.

Sentence A	**Sentence B**	**ROUGE score**	**Semantic similarity**
I love working in this office.	I do not love working in this office.	Very high - .857 (implies high similarity)	Very low
Your plan is a good one.	The idea you propose seems great.	Very low – 0 (implies high dissimilarity)	Very high

Table 5.1 Demonstration of ROUGE weaknesses

Another reason why this thesis cautions against relying on the ROUGE is because it tells us nothing about how well a summary captures the key elements of the larger body of text it is meant to summarize. In this research, the larger body of text would be the article. ROUGE leaves unanswered these key questions: are the most important pieces of information captured in a summary? Are the elements necessary for a decision or assessment included in the summary? Instead, the main usage of ROUGE seems to be in a training setting where model performance can be measured against a labeled training set. In the context of summarization, this entails measuring model-generated summaries against reference summaries. ROUGE would not be useful in a situation where a summarization model is placed into production to support the IC mission because in this type of environment, there is no training set against which to measure performance. Since ROUGE is not designed to measure how well an output summary compares to an input text it would have limited use in a production environment. Some in the NLP research community may argue that ROUGE is still a useful metric despite its shortcomings. While this might be true in limited use cases that require a measure for word overlap, it is difficult to support this position in the context of abstractive summarization. As generative models become more capable, their ability to understand and rephrase ideas will grow, which will make ROUGE less and less useful. Furthermore, the inability to rely on ROUGE to evaluate mission-critical production scenarios where summarization models are used without a labeled training set is another piece of evidence showing the weakness of ROUGE.

Llama3 Is a Useful Evaluator of Summaries

Because of the limitations inherent with the ROUGE metric, a large language model, the Llama3 70 billion parameter model, was used to evaluate summary quality. The use of LLMs to evaluate summary qualities is an area of active contemporary research in the NLP field.[117] A conclusion of this thesis is that Llama3 offers flexibilities and capabilities that would not be possible with ROUGE. An LLM can be instructed to programmatically evaluate multiple bodies of text simultaneously, whereas ROUGE is limited to comparing two bodies of text. For example, in one of the experiments in this thesis, Llama3 was instructed to compare seven bodies of text (the article and six model-generated summaries) and then choose the best summary. Additionally, Llama3 was instructed to run numerous iterations of all the experiments to change the order in which the summaries were presented to it. Ultimately, the use of LLMs in text summarization experiments allows researchers to design specialized and unique experiments that are tailored to specific use cases. While Llama3's findings align with the author's limited human evaluation of Llama2 summaries as the best option, it's important to note that the sample size was small. Consequently, the evidence is insufficient to definitively conclude that Llama3's

[117] Tianyi Zhang et al., "Benchmarking Large Language Models for News Summarization," *Transactions of the Association for Computational Linguistics* 12 (2024): 39-57, https://doi.org/10.1162/tacl_a_0; Fabian Retkowski, "The Current State of Summarization," in Beyond Quantity: Research with Subsymbolic AI (forthcoming, November 2023; published online August 1, 2023), 1, 6-7, https://arxiv.org/abs/2305.04853.0632; Fabian Retkowski, "The Current State of Summarization," in *Beyond Quantity: Research with Subsymbolic AI* (forthcoming, November 2023; published online August 1, 2023), 6-7, https://arxiv.org/abs/2305.04853.

overwhelming preference for Llama2 summaries is warranted. In conclusion, Llama3 was a powerful and flexible evaluator of summaries, but questions remain as to whether an LLM is a reliable evaluator of summary quality, and this is currently an area of active research.[118]

While this thesis recommends using an LLM to evaluate summary quality, it is important to note that not all LLMs are created equal. For example, the Llama2 70 billion parameter model was initially used to evaluate a dataset that contained an article and three summaries. The model was instructed to read the article and then choose the best summary from the three options provided. When an initial five-record sample of the dataset was run, the Llama2 model chose the third summary every time. The order in which the three summaries were provided to Llama2 was then rearranged, for example by presenting what was previously the third summary as the first summary. The results were the same – Llama2 chose the last summary every time, which indicated Llama2 might be showing recency bias in its selection by overestimating the quality of the last summary (the most recent summary it was fed). A larger, 15-record sample of the dataset was then input into Llama2, and for each of the 15 records, Llama2 again selected the last summary every time. At that point, Llama2 was discarded as an evaluator because of its recency bias. When Llama3 was instantiated as an evaluator, it did not show recency bias. Based upon these findings, a recommendation of this thesis is that the Llama2 model is not a suitable evaluator of summarization quality.

[118] Retkowski, "The Current State of Summarization," 4.

Llama2 Performed Better than the Other Models at the Summarization Task

A series of experiments was conducted in which Llama3 was instructed to choose the best summary from a series of candidates. The strongest evidence resulting from these experiments was derived from a test in which Llama3 was instructed to select the best summary from all the model-generated summaries, including the summaries generated by the Llama2 model. Llama3 preferred the Llama2 summaries 98% of the time, showing that the most significant variable in summary quality (as measured by Llama3) was the model that generated it. Llama3's strong preference for Llama2 is noteworthy for three important reasons. First, this finding did not align with the ROUGE scores, which is further evidence that ROUGE is not a very useful metric. Despite Llama3's nearly unanimous selection of the Llama2 summarizations as the best, Llama2's ROUGE scores were among the worst of the six models tested. For example, the Llama2 rouge1 and rougeL scores were the second lowest while the Llama2 rouge2 scores were the lowest. Secondly, there is a possibility that Llama3 may have a natural preference for the style of text generated by Llama2 because both models were built and trained by Meta and are based on the same overarching architecture. While Meta reports that the Llama3 model was trained on a dataset that is seven times as large as the one on which Llama2 was trained, it is possible that both models have other characteristics in common, like training parameters, weights, and biases, that could influence Llama3 to prefer the summaries produced by Llama2.[119] Additionally, it is likely that Llama2 may have seen the CNN/Daily Mail dataset when it was being trained and could therefore perform well on this dataset. Finally, there is a

[119] "Introducing Meta Llama 3: The most capable openly available LLM to date, "Meta, April 18, 2024, https://ai.meta.com/blog/meta-llama-3/.

vast difference in size between the pre-trained language models and the Llama2 LLM. T5 has 60 million parameters, BART has 139 million parameters, Pegasus has 568 million parameters, and Llama2 has 70 billion parameters. In the transformer architecture, model size and scale significantly influence language understanding and generation capabilities. A third and final remark on Llama3's overwhelming preference for Llama2 summaries is that every model except Llama2 produced summaries that were lowercased. Llama2, as the only model that produced summaries with proper capitalization, might have had an unfair advantage. Llama3 could have identified other summaries as inferior due to incorrect casing. While all these may potentially explain why the Llama3 model selected the Llama2 summaries as best in 98% of the cases, experiments to explore this condition further were not part of this research.

As ROUGE scores did not provide a way to validate Llama3's selection of the best summary, redundancy was introduced into the experiment conditions to increase the level of certainty associated with Llama3's selections. Llama3 voted on the summaries it thought was best six times, and for each iteration the order in which the summaries were presented to it changed. The aggregate voting was then tallied. The voting reflected instances in which Llama3's opinion on which summary was the best vacillated. In these cases, Llama3 was inconsistent and might have voted for model A three times and model B three times, which meant there was no winner for that record. However, when Llama3 voted consistently, even when the order of the summaries changed, it demonstrated that Llama3's selections were not capricious or random. Consistent voting added weight to the findings and reduced the ambiguity, and the votes for Llama2 summaries were by far the most consistent at 98%. Despite this strong evidence, it should be noted that the results of the LLM evaluations might have been different if another model of LLM had been used to evaluate the summaries. It is entirely possible that another LLM besides Llama3 may have preferred the BART, T5, or Pegasus summaries over the Llama2 summaries.

It is instructive to compare the Llama2 summaries to those generated by other models. Doing so elucidates the qualitative characteristics of the summaries and provides human insight as to why Llama3 had such a strong preference for the Llama2 summaries. The below three tables compare Llama2 summaries to BART, BART fine-tuned, and T5 fine-tuned summaries.

Article overview	Llama2 summary	BART summary
Air Force member charged with murder of baby	Phoeteama Janiece Neal, a US Air Force mother, has been charged with attempted murder after allegedly trying to drown her 20-month-old daughter when she was denied permission to take her on an overseas deployment. Neal, who was stationed at Luke Air Force Base in Arizona, was going through a custody battle with the child's father and was told that custody would be granted to him while she was deployed to Turkey. After being informed that she could not take her daughter with her, Neal contacted her commander and requested that she not be deployed. When she was told that the Air Force would "get back to her," she became distraught and attempted to	a us air force mom has been charged with attempted murder after she allegedly tried to drown her 20-month old daughter when she was told that she could not take her on an overseas deployment. phoeteama janiece neal, 28, was stationed at luke air force base in glendale, arizona and was being deployed to turkey on august 31. she was also going through a custody battle with the child’s father. mugshot: phooeteama jodi neal has been arrested for attempted murder. the . mom, who had full custody of her daughter up until this month,
Critics of a proposed rail line put forward an alternate route	An alternative route for a high-speed rail project has been proposed by critics of HS2, which would use the former Great Central Line that was closed in the 1960s. The proposal suggests reopening the line, which ran from London to Nottingham with connections to Leeds and Manchester, at a cost of £6billion, significantly lower than the £42.6 billion estimated for the HS2 project. Supporters argue that reopening the Great Central Railway would address environmental concerns and have the support of major retailers eager to transport goods via rail. While	critics of the hs2 rail project have put forward an alternative route using a train line which was closed in the 1960s. the rival scheme would see the reopening of the former great central line, which ran from london to nottingham, with links to leeds and manchester. supporters of the proposal claim it would cost around £6billion, instead of the £42.6 billion which the revamp of the line is set to cost. rival scheme: hs 2 critics have put ahead an alternative routes using

Article overview	Llama2 summary	BART summary
	much of the original infrastructure remains intact, local resident disruption would be minimized, making it a more attract	the former Great central line , which ran between londonto nottedham and was

Table 5.2 Comparison of Llama2 summaries to BART summaries

Article overview	Llama2 summary	BART fine-tuned summary
A roundup of sports stories from across Europe	This article discusses various topics related to European football clubs. In Italy, Juventus Manager Antonio Conte may be leaving the club after failing to advance in the Champions League, while Inter Milan Right-Back Javier Zanetti is set to retire at the end of the season. Roma Owner James Pallotta has arrived in Italy for the Serie A fixture against Juventus. In Spain, Barcelona's Lionel Messi is determined to lead the team to the La Liga title, while Atletico Madrid Striker Diego Costa suffered a hamstring injury. Real Madrid Coach Carlo Ancelloti will decide whether Cristiano Ronaldo, who claims he wants to play, will feature in the match versus C	italian newspapers tuttosport and corriere dello sport both lead with reports that juventus manager antonio conte could be set to leave the club this summer .cds: conte has led juve to three consecutive serie a league titles .
A diamond exhibit at Buckingham Palace for the upcoming Queen’s Diamond Jubilee	The article discusses the upcoming exhibition at Buckingham Palace, which will feature the Cullinan Diamond, the largest diamond ever found, and the jewels created from it. The diamond was discovered in South Africa in 1905 and was later cut into nine large stones and numerous smaller ones. The largest two stones were set into the Sovereign's Sceptre and the Imperial State Crown, while the other seven were mounted into various pieces of jewelry worn by Queen Mary and Queen Elizabeth II. The article provides a brief history of the diamond and the jewelry made from it, and notes that the value of the jewelry is "impossible to calculate"	new: cullinan v brooch, weighing 94.4 carats and 63.3 carats, will be displayed in public for the first time in a unique exhibition at buckingham palace .

Table 5.3 Comparison of Llama2 summaries to BART fine-tuned summaries

Article overview	Llama2 summary	T5 fine-tuned summary
The diary of a British WWII POW was found and returned to his daughter	A British prisoner named Raymond Jones kept a secret diary while incarcerated in Hong Kong's Stanley Camp during World War II, detailing daily life, struggles, and a passionate affair with a fellow internee named Gwen Flower. His wife Marjorie discovered the diary and was heartbroken, tearing up their wedding photos and ordering Raymond to destroy the diary. However, he secretly kept the diary and eventually sent it to a friend or fellow internee, who passed it down to his daughter Alison McEwan. She attempted to locate Raymond's daughter Rae Shaw, who finally received the diary in 2010, 65 years after it	raymond jones kept a meticulous record of daily life in stanley camp . the diary was discovered by alison mcewan, the daughter of another hong kong civil servant .rae shaw, 72, was shocked and heartbroken when she failed to find her .
A career criminal does not have to pay back proceeds from criminal activity	A British man named Declan Madigan, labeled as a "one-boy crime wave," was sentenced to 11 years in prison for a drive-by shooting and drug offenses. He was the youngest person in Britain to receive an ASBO at age 14. Despite over 100 arrests, Madigan claimed that £26,000 in drug money was stolen from him and is not required to pay it back. His lawyer stated that Madigan's relationship with the police is not always a priority for him.	declan madigan was the youngest person in britain to be given asbo in 2000 . 27-year-old was labelled a 'one-boy crime wave' after being arrested 100 times . he was jailed this year for 11 years after being convicted of drive-by shooting .

Table 5.4 Comparison of Llama2 summaries to T5 fine-tuned summaries

The above examples reveal qualitative characteristics of the Llama2 summaries as compared to other model-generated summaries. Generally, the Llama2 summaries read well. The BART model seemed to be mostly extractive, copying the first few lines of the article and changing a few words. The Llama2 summaries tend to capture important concepts and context from later in the article that the other models missed, for example the environmental concerns in the railway article. The BART fine-tuned and T5 fine-tuned summaries tend to be shorter, and thus miss important details. Also, the Llama2 summaries on occasion set useful context by prefacing the summary with "this article discusses." Overall, from a qualitative standpoint and based upon a limited sample size, the Llama2 summaries seem to be better than the other models. It should be noted, however, that human evaluation is somewhat capricious and inconsistent.[120] However, additional evidence to support the finding that Llama2 summaries are better than the other models is the Llama3 voting which showed that the Llama2 summaries were better than the other models nearly every time. This finding aligns with contemporary research which has demonstrated that LLMs outperform more traditional pre-trained language models in the summarization task.[121] Therefore, a conclusion of this research is that the Llama2 summaries appear to be better than the other models, but more extensive testing and evaluation is needed.

Remarks on the Use of LLMs to Generate Summaries within an IC Environment

The Llama3 voting and the qualitative review in this study support a conclusion that the Llama2 summaries are better than the other models that were tested. This conclusion is supported by contemporary research.[122] Although an LLM, like Llama2 or Llama3, would likely generate good summaries in a production environment, the drawbacks of these models should be carefully evaluated by the IC before deployment in mission-critical, national security applications. LLM hallucinations, which are model outputs inconsistent with verifiable facts, pose a significant concern when using these models. In a summarization pipeline, hallucinations can take the form of totally false information or somewhat inaccurate details, such as incorrect

[120] Inderjeet Mani, *Automatic Summarization* (Philadelphia: John Benjamins Publishing Company, 2001), 225.

[121] Xiao Pu, Mingqi Gao, and Xiaojun Wan (Wangxuan Institute of Computer Technology, Peking University), "Summarization is (Almost) Dead," working paper (Ithaca: Cornell University Library, arXiv.org, September, 2023): 1-2, 4. https://doi.org/10.48550/arXiv.2309.09558.

[122] Pu, Gao, Wan, "Summarization is (Almost) Dead," 1.

dates, times, or names.[123] There is currently no consensus on an effective solution to LLM hallucinations.[124] LLM hallucinations may seem true because of the confidence and fluency of the LLM's statements. A naïve or inexperienced user may be lulled into believing the model's summary even though it contains inaccurate information. While hallucinations might be bothersome or detrimental in a business setting, the stakes are higher in an IC environment. For example, if an LLM includes inaccurate information in a summary of fast-breaking events with national security implications, policymakers that rely on the summary may make misguided decisions. The IC can offset the risk of hallucinations in summarization pipelines by keeping a human-in-the-loop, which would prevent the wholehearted reliance on model-generated summaries.[125] However, this comes with cost implications and reduces the speed at which these models can support national security decision making or intelligence analysis. Another consideration for the IC is the technical challenges associated with deploying an LLM in a secure, air-gapped, classified environment. For this use case, substantial compute resources will be required for real-time inference at scale and potential model training.[126] While challenging,

[123] Lei Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," (Ithaca: Cornell University Library, arXiv.org, November 9, 2023): 1, https://arxiv.org/abs/2311.05232.

[124] Lei Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," 31.

[125] Maryam Amirizanianiet al., "LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop," (Ithaca: Cornell University Library, arXiv.org, May 22, 2024): 1-2 https://arxiv.org/abs/2402.09346.

[126] Suhas Pai, *Designing Large Language Model Applications: A Holistic Approach* (forthcoming, January 2025; available online at O'Reilly Media. Inc), chapter 4, "Adapting LLMs to Your use Case," https://learning-oreilly-com.library.access.arlingtonva.us/library/view/designing-large-language/9781098150495/.

several organizations and groups within the IC have already solved this problem and can provide guidance to other IC organizations that are seeking to implement these models.

As an alternative, IC organizations can choose to rely on smaller pre-trained language models for summarization tasks, such as BART, T5 or Pegasus. These models can easily be brought into an environment with a few lines of code in a Jupyter Lab notebook and can even be run on a local machine.[127] The smaller models are not known to suffer from hallucinations as severely as large language models. However, using them involves several tradeoffs. One of the most important is summary quality. Based upon the preliminary findings of this study as well as other contemporary research, the smaller models are not as good at summarizing text as the LLMs.[128] Compared to LLMs, models like BART, T5, and Pegasus generally have smaller token limits for both input and output. Consequently, truncation may be required for lengthy input text, and the output text might fall short of meeting the requirements for a given use case. In conclusion, the smaller models are a good choice for experimentation and limited-use cases within the IC. An LLM is a more suitable model for summarization tasks at scale.

Measuring Summarization Quality is Difficult, but Several Promising Approaches Exist

A key finding of this research is that evaluating summary quality is inherently difficult. Summaries are generated in many different contexts with a countless number of variables influencing how they are produced, used, consumed, and evaluated. Summaries serve different

[127] Sinan Ozdemir, "Hands-on T5" in "Introduction to Transformer Models for NLP: Using BERT, GPT, and More to Solve Modern Natural Language Processing Tasks," (O'Reilly Media: February 2023), https://learning-oreilly-com.library.access.arlingtonva.us/videos/introduction-to-transformer/9780137923717/9780137923717-TMN1_01_11_01/.

[128] Xiao Pu, Mingqi Gao, and Xiaojun Wan (Wangxuan Institute of Computer Technology, Peking University), "Summarization is (Almost) Dead," working paper (Ithaca: Cornell University Library, arXiv.org, September, 2023): 1-2, 4. https://doi.org/10.48550/arXiv.2309.09558.

purposes in different situations. In the IC, a summary of a body of intelligence reporting may be used to justify a recommendation for military or covert action. On the other hand, a summary of a meeting might serve to inform team members of important developments. There are nearly limitless variations in what a summary is and what purpose it serves. There is simply no such thing as an ideal or perfect summary. Despite these challenges, there are frameworks and approaches that are useful for evaluating summaries. One of these is identifying the key points that must be included in a summary for a particular use case, then testing whether the model-generated summary includes the key points. For example, if a situation requires a recipient to follow a set of steps to accomplish a task, then the summary would have to include the necessary steps. If a situation calls for a decision to be made based upon a given set of circumstances and conditions, for example increasing security posture at a facility, then a summary would have to include all the relevant conditions that enable the decision. This type of approach to evaluating summarization, based upon intent of the summary, could lead to better metrics and better models. It entails being able to identify the most important pieces of information for a use case, then measuring whether the presence or absence of the key information in the summary enables or prevents the intended use case.

Organizational Issues: Culture, Resources, and Data

This research also revealed several strategic, organizational-level concepts that are relevant for IC organizations that are seeking to implement summarization models in production. The most important one of these is culture.[129] It is axiomatic that an organizational culture that

[129] Brian Katz, "Maintaining the Intelligence Edge: Reimagining and Reinventing Intelligence through Innovation," (Center for Strategic & International Studies (CSIS), Jan 2021), ix.

embraces and encourages innovation is a prerequisite for deploying ML models at scale. However, there is a key corollary that organizations may miss. It is not enough to write a strategy that says AI is important, or to talk about the need for innovation, or to proclaim that decision-making should be data-driven. Any organization can do that. A critical ingredient in whether an organization achieves the technological advancements necessary to benefit from AI comes down to whether sufficient resources are allocated to support the AI strategy and aspirations.[130] This means spending time and money on talent, training, compute infrastructure, and process improvements. It also means that other perceived priorities will need to be de-emphasized. In the IC, when nearly everything sems like a priority and budget increases for new, unproven technologies are hard to secure, it is no surprise that AI and ML efforts languish. For summarization, at a minimum, an organization will need a compute environment – a secure, cloud-based infrastructure that data scientists, ML/AI engineers, and even curious practitioners can easily access without cost, limitations, restrictions, or bureaucratic roadblocks. The compute environment is an enabler of innovation. For summarization tasks that involve even the most basic model training experiments, it is crucial for the compute environment to be GPU-enabled.[131] The experiments for this thesis ran into serious compute challenges caused by GPU limitations. It was impossible to fine-tune BART and T5 on the basic web-based NIU Jupyter Lab compute environment. As a workaround, access was provided to a more powerful GPU-enabled NIU compute environment with 24GB of GPU RAM. This environment was not strong

[130] Katz, "Maintaining the Intelligence Edge: Reimagining and Reinventing Intelligence through Innovation," xi.

[131] Hongyu Zhu et al. "TBD: Benchmarking and Analyzing Deep Neural Network Training," (Ithaca: Cornell University Library, arXiv.org, April 14, 2018): 1-2, https://arxiv.org/abs/1803.06905.

enough to fine-tune the BART and T5 models, nor could it support more than a few users running GPU-intensive training loops simultaneously. To circumvent these limitations, PNNL graciously provided access to a high-performance dual GPU system with 96GB of GPU RAM located at the PNNL iRES facility in Washington, D.C. For any serious mission-critical AI/ML research and experimentation, IC organizations must understand the need for expansive GPU compute capacity and find ways to provide the necessary resources. Otherwise, great ideas may stall before they even get started.

This research also led to several findings about data that are relevant to strategic leaders of IC organizations. Finding a sufficiently large, high-quality, domain-specific dataset for this research was challenging. A dataset with text and human-generated summaries was needed for this research to compare model-generated summaries with human ones. One such dataset was identified, but organizational policies prevented its use for this project. Because data are so critical for AI/ML experiments, IC leaders should understand the value of making high-quality datasets available.[132] By making data discovery easy, data scientists and AI/ML engineers within the IC can use these datasets to derive new insights, discover new patterns, or enhance national-security decision making. IC leaders can enable AI/ML efforts within their organizations by conducting user research to understand the critical data needs of an organization from the AI/ML perspective. This research could then be used as the basis for building data libraries and data repositories that are fresh, reliable, trustworthy, complete, secure, and easy for users to access.

[132] Katz, “Maintaining the Intelligence Edge: Reimagining and Reinventing Intelligence through Innovation,” 29.

Recommendations for Future Research

This thesis identified several exciting areas for future research that could advance the body of knowledge in the text summarization field. One of these is research on LLM hallucinations.[133] This research might entail scientific or mathematical testing to determine whether a summary contains information that is not supported by the input text. Another way to approach the hallucination problem in the context of text summarization would be to examine how retrieval-augmented generation (RAG) could be used to establish authoritative sources of information which LLMs must use to generate reliable information.[134] This could involve testing the extent to which the use of RAG controls LLM hallucinations. Additional research pathways to address the phenomenon of LLM hallucinations include object hallucinations in large vision-language models; using internal error-correction mechanisms to mitigate hallucinations in LLMs without dependence on external knowledge sources; and studies to understand the boundaries of LLM knowledge and quantify uncertainty in LLM outputs.[135]

More research is needed on the explainability of LLMs, sometimes referred to as the black box problem. This continues to be a vexing problem with ethical implications.[136] How a

---

[133] Lei Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," (Ithaca: Cornell University Library, arXiv.org, November 9, 2023): 1-8, https://arxiv.org/abs/2311.05232.

[134] Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," 25.

[135] Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," 29-31.

[136] Haoyan Luo and Lucia Specia, "From Understanding to Utilization: A Survey on Explainability for Large Language Models," (Ithaca: Cornell University Library, arXiv.org, February 22, 2024): 1, https://doi.org/10.48550/arXiv.2401.12874; Xuansheng Wu et al., "Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era," (Ithaca: Cornell University Library, arXiv.org, Cornell University Library, arXiv.org, March 13, 2024): 1-5, https://doi.org/10.48550/arXiv.2403.08946.

model decides what to include in a summary and whether the model can tell us where the information comes from is an important consideration, especially in the IC domain. This is the question of attribution.[137] Determining the provenance of the information in a summary is important for IC decision-making. Without this assurance, analysts might be unable to identify the source of information which raises ethical concerns if national security decisions are based upon these products, especially if they involve using force. The IC will benefit from more studies on LLM explainability, and this is an active area of research.[138]

Research that leads to reliable, automated metrics to replace ROUGE would be very useful.[139] Potential avenues of research on automated metrics might involve dimensionality reduction of the multi-dimensional vector space embeddings to enable principal component analysis that measures semantic similarity between the summary and input text. Clustering algorithms might also be helpful in this regard. Research on automated metrics might include methods to account for the presence or absence of vital pieces of information in a model-generated summary. Research that examines quantitative methodologies for human evaluation of summaries could also be helpful for the development of new metrics. This would involve the development of rubrics, scales, and parameters that provide ground-truth validation of proposed

[137] Xuansheng Wu et al., "Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era," 6-8.

[138] Haoyan Luo and Lucia Specia, "From Understanding to Utilization: A Survey on Explainability for Large Language Models," 9.

[139] Alexander R. Fabbri et al., "SummEval: Re-Evaluating Summarization Evaluation," *Transactions of the Association for Computational Linguistics* 9 (2021): 401, https://doi.org/10.1162/tacl_a_00373; Vivek Srivasta, Savita Bhat, and Niranjan Pedanekar, "Hiding in Plain Sight: Insights into Abstractive Text Summarization," *The Fourth Workshop on Insights from Negative Results in NLP, Association of Computational Linguistics* (2023): 68, http://dx.doi.org/10.18653/v1/2023.insights-1.8.

automated metrics. Another research approach that would support automated metrics would be to use a panel of LLMs to evaluate summarization quality.[140] Whereas this study used a single LLM, Llama3, as an evaluator, the panel approach would employ a diverse set of LLMs to evaluate summary quality. This approach might lead to robust and reliable metrics to measure the performance of summarization models.[141]

Two additional areas of research are recommended for text summarization: multi-document summarization and summarization research that deals with a national security corpus of text.[142] Research into multi-document summarization is important because it would address the practical need of condensing a body of emails, reports, or other documents into a single summary. This is a common task within the IC. Ideally, research into multi-document summarization would lead to work surfaces that enable drag and drop functionality, making multi-document summarization as easy as attaching files to an email. Summarization research using a high-quality corpus of intelligence-related text would reveal the extent to which language models can generalize to the national security domain.

---

[140] Pat Verga et al., "Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models," (Ithaca: Cornell University Library, arXiv.org, April 29, 2024): 1, https://doi.org/10.48550/arXiv.2404.18796.

[141] Verga et al., "Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models," 2.

[142] Fabian Retkowski, "The Current State of Summarization," in *Beyond Quantity: Research with Subsymbolic AI* (forthcoming, November 2023; published online August 1, 2023), 7, https://arxiv.org/abs/2305.04853.

APPENDIX

Code and Notebooks

The code and notebooks can be found at https://code.pnnl.gov/collective-intelligence/developer-spaces/mlevine/levine_thesis.

BIBLIOGRAPHY

Amirizaniani, Maryam, Jihan Yao, Adrian Lavergne, Snell Okada Elizabeth, Aman Chadha, Tanya Roosta, and Chirag Shah. “LLMAuditor: A Framework for Auditing Large Language Models using Human-in-the-Loop.” Ithaca: Cornell University Library, arXiv.org, (May 22, 2024): 1-14. https://arxiv.org/abs/2402.09346.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate.” Conference paper at the *2015 International Conference of Learning Representations* (ICLR), 2015. https://arxiv.org/abs/1409.0473.

Chen, Tong, Xuewei Wang, Tianwei Yue, Xiaoyu Bai, Cindy X. Le, and Wenping Wang. “Enhancing Abstractive Summarization with Extracted Knowledge Graphs and Multi-Source Transformers.” *Applied Sciences* 13, no. 13: 7753 (June 2023): 1-14. https://doi.org/10.3390/app13137753

Cullingford, Richard Edward. “Script Application: Computer Understanding of Newspaper Stories.” Master's thesis, Department of Computer Science, Yale University , Jan. 1978. In PROQUESTMS ProQuest Dissertations & Theses Global, https://niu.idm.oclc.org/login?url=https://www.proquest.com/dissertations-theses/script-application-computer-understanding/docview/288085255/se-2.

Dahl, George E., Dong Yu, Li Deng, Alex Acero. “Large Vocabulary Continuous Speech Recognition with Context-Dependent DBN-HMMS.” *2011 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (2011): 4688 - 4691, https://doi.org/10.1109/ICASSP.2011.5947401.

Dahm, Michael. “Chinese Debates on the Military Utility of Artificial Intelligence.” *War on the Rocks* (website), June 5, 2020. https://warontherocks.com/2020/06/chinese-debates-on-the-military-utility-of-artificial-intelligence/.

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, https://doi.org/10.1162/tacl_a_00373.

Fregly, Chris, Antje Barth, Shelbee Eigenbrode. *Generative AI on AWS: Building Context-Aware Multimodal Reasoning Applications*. Sebastopol, CA: O'Reilly Media, Inc., December, 2023.

Géron, Aurélien. “Natural Language Processing Using Transformer Architectures.” Recorded October, 2019. Tensor Flow World. https://learning-oreilly-com.library.access.arlingtonva.us/videos/natural-language-processing/0636920373605/0636920373605-video329383/.

Giarelis, Nikolaos, Charalampos Mastrokostas, and Nikos Karacapilidis. "Abstractive vs. Extractive Summarization: An Experimental Review." Applied Sciences 13, no. 13: 7620 (June 2023): 1-20. https://doi.org/10.3390/app13137620.

Graesser, Laura and Wah Loon Keng. *Foundations of Deep Reinforcement Learning: Theory and Practice in Python*. Boston: Addison Wesley, 2020.

Hou, Sheng-Luan, Xi-Kun Huang, Chao-Qun Fei, Shu-Han Zhang, Yang-Yang Li, Qi-Lin Sun and Chuan-Qing Wang. "A Survey of Text Summarization Approaches Based on Deep Learning." *Journal of Computer Science and Technology* 36, no. 3 (May 31, 2021): 633-663. http://dx.doi.org.niu.idm.oclc.org/10.1007/s11390-020-0207-x.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions." Ithaca: Cornell University Library, arXiv.org, (November 9, 2023): 1-49. https://arxiv.org/abs/2311.05232.

Intelligence Community Chief Data Officer. *The IC Data Strategy: 2023-2025*. https://www.dni.gov/files/ODNI/documents/IC-Data-Strategy-2023-2025.pdf.

Katz, Brian. "Maintaining the Intelligence Edge: Reimagining and Reinventing Intelligence through Innovation." Center for Strategic & International Studies (CSIS). (Jan 2021). https://www.csis.org/analysis/maintaining-intelligence-edge-reimagining-and-reinventing-intelligence-through-innovation.

Krizhevsky, Alex, Ilya Sutskever, Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, edited by: F. Pereira, C.J. Burges, L. Bottou and K.Q. Weinberger. Red Hook, NY: Curran Associates, Inc., 2013. https://papers.nips.cc/paper_files/paper/2012.

Krohn, Jon, Grant Beyleveld, and Aglae Bassens. *Deep Learning Illustrated*. Boston: Addison Wesley, 2020.

Lehnert, Wendy G. "Affect Units and Narrative Summarization," research report #179, Department of Computer Science, Yale University (May 1980). Defense Technical Information Center (DTIC). https://apps.dtic.mil/sti/citations/ADA086735.

_____. "Narrative Text Summarization." Proceedings of the AAAI Conference on Artificial Intelligence, 1, (1980): 337-339. https://aaai.org/papers/00337-narrative-text-summarization/.

Lehnert, Wendy G. and Mark H. Burstein. "The Role of Object Primitives in Natural Language Processing," research report #162, Department of Computer Science, Yale University

(January 1979). Defense Technical Information Center (DTIC). https://apps.dtic.mil/sti/citations/ADA069861.

Lin, Chin-Yew. "ROUGE: A Package for Automatic Evaluation of Summaries." In *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics (2004): 1-8. https://aclanthology.org/W04-1013.

Luhn, H.P. "The Automatic Creation of Literature Abstracts." In *IBM Journal of Research and Development* 2, issue 2 (April, 1958): 159-165. https://doi.org/10.1147/rd.22.0159.

Luo, Haoyan and Lucia Specia. "From Understanding to Utilization: A Survey on Explainability for Large Language Models." Ithaca: Cornell University Library, arXiv.org (February 22, 2024): 1-13. https://doi.org/10.48550/arXiv.2401.12874.

Makhoul, John, Le Zhang Sanjay, Krishna Gouda, Rich Schwartz, William Hartmann, Lee Tarlin, Damianos Karakos, Manaj Srivastava, David Akodes. "Foreign Language Automated Information Retrieval (FLAIR)/Machine Translation for English Retrieval Of Information In Any Language (MATERIAL)," report number AFRL-RH-WP-TR-2021-0088, Air Force Research Laboratory (December 2021). Defense Technical Information Center (DTIC). https://apps.dtic.mil/sti/trecms/pdf/AD1165370.pdf.

Mani, Inderjeet. *Automatic Summarization*. Philadelphia: John Benjamins Publishing Company, 2001.

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. "On Faithfulness and Factuality in Abstractive Summarization." Working paper. Ithaca: Cornell University Library, arXiv.org, (May 2020): 1-14. http://arxiv.org/abs/2005.00661.

McKeown, Kathleen et al. "System for Cross-Language Information Processing, Translation and Summarization (SCRIPTS)," report number AFRL-RH-WP-TR-2021-0113, Air Force Research Laboratory (December 2021). Defense Technical Information Center (DTIC). https://apps.dtic.mil/sti/trecms/pdf/AD1165721.pdf.

Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. "Large Language Models: A Survey." Ithaca: Cornell University Library, arXiv.org, (Feb. 2024): 1-43. https://arxiv.org/html/2402.06196v2.

Office of the Director of National Intelligence. *2023 National Intelligence Strategy*. https://www.dni.gov/files/ODNI/documents/National_Intelligence_Strategy_2023.pdf.

Ozdemir, Sinan. "Introduction to Transformer Models for NLP: Using BERT, GPT, and More to Solve Modern Natural Language Processing Tasks." Recorded February, 2023. O'Reilly training video. https://learning-oreilly-com.library.access.arlingtonva.us/course/introduction-to-transformer/9780137923717/.

_____. *Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs*. Hoboken, NJ: Addison Wesley, 2024.

Pu, Xiao, Mingqi Gao, and Xiaojun Wan (Wangxuan Institute of Computer Technology, Peking University). "Summarization is (Almost) Dead." Working paper. Ithaca: Cornell University Library, arXiv.org, (September, 2023): 1-9. https://doi.org/10.48550/arXiv.2309.09558.

Rao, Delip and Brian McMahan. *Natural Language Processing with PyTorch*. Sebastopol, CA: O'Reilly Media, Inc., 2019.

Ravi, Janani. "AI Text Summarization with Hugging Face." Recorded October, 30 2023. LinkedIn Learning video. https://www.linkedin.com/learning/ai-text-summarization-with-hugging-face/evaluation-metrics-for-summaries.

Retkowski, Fabian. "The Current State of Summarization." In *Beyond Quantity: Research with Subsymbolic AI* (forthcoming, November 2023; published online August 1, 2023). http://arxiv.org/abs/2305.04853

Rieger, Charles J., III. "Conceptual Memory: A Theory and Computer Program for Processing the Meaning Content of Natural Language Utterances." PhD diss., Department of Computer Science, Stanford University , Apr. 1974. ProQuest Dissertations & Theses Global, https://niu.idm.oclc.org/login?url=https://www.proquest.com/dissertations-theses/conceptual-memory-theory-computer-program/docview/302726268/se-2?accountid=10504.

Rush, Alexander M., Sumit Chopra, and Jason Weston. "A Neural Attention Model for Abstractive Sentence Summarization." In *Conference Proceedings: Conference on Empirical Methods in Natural Language Processing* (EMNLP), 2015, Lisbon. Red Hook, NY: Curran Associates, Inc., (2015): 379-389.

Singh, Jyotika. *Natural Language Processing in the Real World : Text Processing, Analytics, and Classification*. Boca Raton, FL: CRC Press LLC, 2023. ProQuest eBook Central, https://ebookcentral.proquest.com/lib/niulibrary-ebooks/detail.action?docID=7250810.

Srivastava, Vivek, Savita Bhat, Niranjan Pedanekar. "Hiding in Plain Sight: Insights into Abstractive Text Summarization." *The Fourth Workshop on Insights from Negative Results in NLP, Association of Computational Linguistics* (2023): 67–74. http://dx.doi.org/10.18653/v1/2023.insights-1.8.

Tunstall, Lewis, Leandro Von Werra, Thomas Wolf. *Natural Language processing with Transformers, Revised Edition: Building Language Applications with Hugging Face*. Sebastopol, CA: O'Reilly Media, Inc., May 2022.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is All You Need." In *NIPS '17: Proceedings of the Annual Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc., (December 2017): 6000-6010. https://dl.acm.org/doi/10.5555/3295222.3295349.

Verga, Pat, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. "Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models." Ithaca: Cornell University Library, arXiv.org (April 29, 2024): 1-17. https://doi.org/10.48550/arXiv.2404.18796.

The White House. *National Security Strategy*. October 2022. https://www.whitehouse.gov/wp-content/uploads/2022/10/Biden-Harris-Administrations-National-Security-Strategy-10.2022.pdf.

Winograd, Terry. "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language." Revised version of PhD diss., Massachusetts Institute of Technology, Jan. 1971. https://dspace.mit.edu/handle/1721.1/15546.

Wu, Xuansheng, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, and Ninghao Liu. "Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era." Ithaca: Cornell University Library, arXiv.org, (March 13, 2024): 1-56. https://doi.org/10.48550/arXiv.2403.08946.

Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. "Benchmarking Large Language Models for News Summarization." *Transactions of the Association for Computational Linguistics* 12 (2024): 39-57. https://doi.org/10.1162/tacl_a_00632.

Zhu, Hongyu, Mohamed Akrout, Bojian Zheng, Andrew Pelegris, Amar Phanishayee, Bianca Schroeder, and Gennady Pekhimenko. "TBD: Benchmarking and Analyzing Deep Neural Network Training." Ithaca: Cornell University Library, arXiv.org, (April 14, 2018): 1-36. https://arxiv.org/abs/1803.06905.