Explainability Meets Text Summarization: A Survey

Mahdi Dhaini, Ege Erdogan, Smarth Bakshi and **Gjergji Kasneci**
School for Computation, Information and Technology
Technical University of Munich, Germany
{firstname.lastname}@tum.de

Abstract

Summarizing long pieces of text is a principal task in natural language processing with Machine Learning-based text generation models such as Large Language Models (LLM) being particularly suited to it. Yet these models are often used as black-boxes, making them hard to interpret and debug. This has led to calls by practitioners and regulatory bodies to improve the explainability of such models as they find ever more practical use. In this survey, we present a dual-perspective review of the intersection between explainability and summarization by reviewing the current state of explainable text summarization and also highlighting how summarization techniques are effectively employed to improve explanations.

1 Introduction

Against the ever-growing influx of textual content, being able to effectively summarize long pieces of text is crucial to extract useful information. Whereas once a significant amount of manual labour would be necessary, now *automatic text summarization* (ATS) can be performed by deep learning models, especially as they grow in capabilities and become more easily accessible (Bubeck et al., 2023). Nevertheless, such deep learning models are essentially black boxes. They provide no immediate information regarding their internals, and they can fail in ways imperceptible to a novice, e.g. by producing incorrect output that looks legitimate and create an illusion of understanding (Messeri and Crockett, 2024; Li, 2023). It is thus of critical importance that such models can be made *explainable*, especially in sensitive fields such as law (Magesh et al., 2024) and healthcare (Mamalakis et al., 2024). In this work, we bridge the gap between text summarization and explainability and highlight through a literature review their dualistic relation, namely that on one side summarization methods help develop explainable methods, and on the other explainability methods help enhance and understand summarization methods. Explainability in summarization can take two forms, each targeting different stakeholders. The first form involves explaining the output of summarization models, intended for the end users of summarization systems. The second form is focused on understanding and interpreting the internal workings and mechanisms of the summarization model, primarily aimed at debugging the model, which is intended for model developers.

Why Text Summarization and Explainable AI(XAI)? An explanation is an attempt at extracting useful, concise information from a complex, black-box model. Likewise a summary attempts to extract the essential bits of a longer piece of text. Seen this way, an explanation *summarizes* the model's prediction, and a summary *explains* the summarized piece of text. It is thus beneficial to consider the two problems together since approaches to one can inform the approaches to the other, as we will provide examples throughout the survey.

Contributions As far as we know, this work is the first to present an overview of explainable text summarization and to offer a dual perspective on how explainability and summarization can mutually contribute to each other. In the scope of this work, we use the terms related to explainability and interpretability interchangeably.

The contributions of this survey are summarized as follows:

- We review the current state of research on the intersection between explainability and text summarization. Our approach is twofold: we explore how explainability is applied to text summarization and how text summarization is utilized to enhance explainability.
- We present an overview and categorization of the explainability techniques and explanations

Proceedings of the 17th International Natural Language Generation Conference, pages 631–645
September 23–27, 2024.

for text summarization.

- We outline the three most used visualization and evaluation approaches for the explanations for text summarization.
- We discuss and draw conclusions on the practical usefulness of explainability approaches in text summarization.
- We highlight the popular models, datasets, and evaluation metrics for text summarization in the reviewed papers.

2 Background

Problem Description. Text summarization is an important problem in NLP around creating short and informative summaries of longer pieces of text. Approaches to text summarization can be in two types: *Abstractive summarization* methods generate new sentences by processing the input sentences (i.e. summarize in their own words), while *extractive summarization* approaches directly copy parts of the input text to construct a summary.

Models. With the development of the transformer architecture (Vaswani et al., 2017), transformer-based models such as T5 (Raffel et al., 2020) are commonly used for text summarization as in many language generation tasks. Summarization can also often benefit from other sources of domain knowledge, such as in knowledge graphs. To enable the use of these different modalities, architectures such as graph neural networks (Kipf and Welling, 2016; Veličković et al., 2018) can also find use in summarization pipelines.

Evaluation. Various metrics can be used to evaluate generated summaries (see Table 4 in the Appendix). The most frequently used metrics are variants of the ROUGE score, in which n-gram overlap between the input and summary texts is measured.

Tailoring summaries to user intents. Summaries can also be tailored to specific user intents, which is particularly challenging when dealing with long-tail user intents. This difficulty arises because even some of the most advanced LLMs today struggle to accurately recognize and address niche intents, as analyzed and discussed by Bodonhelyi et al. (2024). The assessment of intent-driven summarization holds significant potential for further research and novel specialized metrics, capturing the semantic adequacy of a summary and user satisfaction.

3 Methodology

In this survey, we employ a systematic review approach following the methodology defined by Kitchenham and Charters (2007). We detail the review methodology in Appendix A. We first formulated our research questions with a high degree of specificity as follows:

RQ1: What are the popular models, datasets, and evaluation metrics used in existing research on explainable text summarization?

RQ2: What XAI techniques are employed for text summarization in the existing research studies?

RQ3: How are such explanations visualized and evaluated?

RQ4: Can we derive practical conclusions on the usefulness of Explainability techniques for text summarization?

RQ5: How can text summarization methods be utilized by XAI to provide explanations?

We defined a set of related keywords to search for relevant papers and applied the following search string to the title, abstract, and keywords: *("explainable" OR "interpretable" OR "explainability" OR "interpretability") AND ("text summarization")*. We then filter and divide the papers into two categories: (1) *explainability for text summarization* direction, in which explainability techniques are applied to explain the summarization models outputs or internal mechanisms, (2) *summarization for explainability* direction, which consists of papers where text summarization is used to provide explanations independent of the NLP task under consideration.

4 Results

In this section, we present the results of our review, structured according to the research questions formulated earlier and also provide some insights at the end of each section.

4.1 Text Summarization

This section presents the summarization models, evaluation metrics, and datasets used in the studies we reviewed, specifically those where explainability is applied to text summarization. Our aim is not to exhaustively cover all text summarization models, datasets, and metrics but rather to focus on those utilized in the reviewed studies.

4.1.1 Models and Metrics for Text Summarization (RQ1)

Unlike extractive summarization, abstractive summarization approaches involve understanding the underlying semantics of the textual content and generating a new summary that is textually different from the original text. These approaches utilize complex neural network-based models that are black-box models due to their opacity and lack of interpretability. Therefore, explainability techniques are explored for abstractive summarization to ensure end-users understand and trust the summary generation process. This is evident in our results in Table 2, where explainability techniques are mostly applied to abstractive summarization.

While exploring the papers, we noticed that a variety of Pre-trained Langauge Models (PLMs) have been used for the task of text summarization. As shown in Table 2, the most commonly used models include RNNs, GAMs-based models (Hastie and Tibshirani, 1985), and Transformer models, out of which Transformer models, specifically BERT and T5, are the most used ones.

Additionally, GAM-based models have been employed in explainable ATS by da Silva et al. (2023), where they leverage the inherent interpretability of GAMI for extractive ATS. They apply two GAMI-based models, Explainable Boosting Machine (Lou et al., 2013) and GAMI-Net (Yang et al., 2021), as the decision algorithms for summarization. Although the performance of such methods falls short compared to more recent back-box architectures, they provide transparency in the prediction-making process, which is important in extractive ATS. More recently, Xie et al. (2024) propose a novel transformer-based architecture for explainable biomedical extractive summarization by integrating graph neural topic models and domain knowledge into PLMs to enhance performance and explainability.

Insights: we note the lack of information that would allow for reproduction of results, as some works only mention the model types such as *seq2seq* and *transformers* (Wang et al., 2020). Table 2, also reveals the dominance of transformer-based models for explainable text summarization compared to classical seq2seq models (e.g., RNNs, LSTMs). This aligns with our expectations within the scope of this work, given the better performance and less interpretability of transformer-based models.

Table 1: How many times each summary evaluation method was used in the reviewed papers (BES: BERTScore, BAS: BARTScore)

	ROUGE	BES	BAS	BLEU	Human Eval
#	11	1	1	1	7

Evaluating summaries is one of the most critical tasks in ascertaining the quality of generated summaries. Table 1 displays how many times each metric was used to evaluate summaries in the reviewed papers. The ROUGE score is the most extensively used. On a positive note, 7/17 of the papers perform some form of human evaluation, while BERT/BARTScore and BLEU metrics are also used.

4.1.2 Datasets for Text Summarization (RQ1)

Text summarization datasets typically consist of pairs of source documents and their corresponding reference summaries, covering domains such as news articles, scientific papers, Wikipedia articles. Large-scale datasets, such as the CNN/Daily Mail dataset and the New York Times Annotated Corpus, provide diverse and extensive sources for training abstractive and extractive summarization models.

Among the datasets we observed during our survey as mentioned in Table 5 in the Appendix, the CNN/DailyMail dataset is the most frequently used for text summarization. In particular for *explainable* text summarization, Kim et al. (2023) provide the ExplainMeetSum dataset containing meeting summaries with 'ground truth' human-annotated explanation sentences for each summary. Nevertheless, there is a lack of such explainable summarization datasets.

Insights: There is a large literature on text summarization datasets, yet little attention has been paid to curating *explainable* text summarization datasets, e.g., with ground truth explanations. Extending this line of work to different settings can be valuable for developing more faithful summarization methods.

4.2 Explainability for Text Summarization

In this section, we report the results related to explainability for text summarization based on the studies we reviewed.

4.2.1 Categorization of Explanations (RQ2)

In categorizing the generated explanations, we employ two primary criteria. The first criterion clas-

Table 2: Overview of summarization approach, models used across the surveyed papers. HGAT: hierarchical graph attention network. LSA: latent semantic analysis. GAM: Generalized Additive Model. *Authors don't provide additional information on the model(s) used.

Approach (#)	Model	#	References
Abstractive *(12)*	Seq2Seq* HGAT (Zhan et al., 2022)	2	(Wang et al., 2020; Moody et al., 2022)
	BART-Large (Lewis et al., 2020)	2	(Jiang et al., 2024; Wang et al., 2023b)
	T5 (Raffel et al., 2020)	2	(Hongwimol et al., 2021; Ismail et al., 2023)
	Transformers* (Vaswani et al., 2017)	3	(Li et al., 2021; Wang et al., 2021; Kryściński et al., 2020)
	Pointer generator network (See et al., 2017)	1	(Norkute et al., 2021)
	RNN (Elman, 1990)	1	(Majumder et al., 2022)
	PEGASUS (Zhang et al., 2020a)	1	(Saha et al., 2023)
Extractive *(8)*	TextRank and LSA (Mihalcea and Tarau, 2004) BERTSum (Liu and Lapata, 2019) Sentence-BERT (Reimers and Gurevych, 2019) Graph neural networks (Kipf and Welling, 2016; Veličković et al., 2018)	4	(Moody et al., 2022) (Li et al., 2022) (Schaper et al., 2022) (Xie et al., 2024)
	Transformers* (Vaswani et al., 2017)	1	(Li et al., 2021)
	GAM-based models (Hastie and Tibshirani, 1985)	1	(Silva et al., 2022)
	Bi-LSTM (Graves et al., 2013)	2	(Vo et al., 2024) (Reunamo et al., 2022)

sifies explanations based on their scope: *local explanations* are specific to a single prediction for a particular input, while *global explanations* refer to the overall prediction process of the model without being concerned about a specific input. In the reviewed studies, 17 proposed methods out of 19 fall under local explanations, while only two belong to the global explanation category.

The second criterion categorizes methods based on whether they are part of the prediction process or whether they require post-processing after the model's prediction: *self-explaining*, also called *ante-hoc*, refers to explanations presented inherently within the prediction process, such as decision trees, rule-based models, and attention. This category also includes explainability mechanisms that can be integrated during the model's processing phase to provide insights before the final prediction is made, such as injecting interpretable patterns into attention matrices. On the other hand, *post-hoc explanations* require further operation after the prediction process such as LIME (Ribeiro et al., 2016). In the reviewed papers, 10 methods fit the self-explaining category while 9 are considered post-hoc explainability methods. Explainability methods can also be categorized as model-agnostic and model-specific. Post-hoc methods are model-agnostic because they are applied after training, regardless of model type, while self-explainable ones are model-specific as they inherently offer explainability.

Insights: The significantly higher use of local explanations rather than global signals the hardness of obtaining general information about the decision-making process especially for a text generation task compared to e.g. tabular data classification. Local explanations on the other hand provide immediate information about how the current summary was generated.

4.2.2 Categorization of Explainability Techniques (RQ2)

We classify the explainability techniques into four different categories on the basis of the approach they adopt to generate explanations or justifications for the output generated by a black-box model.

Example-driven. These methods discover and show other examples that are semantically comparable to the input instance, usually from available labeled data, in order to explain the prediction of the input instance. It is also an intuitive approach that helps the user gain faith in the predictions being generated. This approach has been utilized in (Wang et al., 2020), where the reviews are summarized in the form of a textual summary and a structured graph. Here, for explainaing the review summaries, a text instance is picked from the original text corpus to explain the generated summary. Ismail et al. (2023) use the Input Reduction (Feng and Boyd-Graber, 2019) and HotFlip

Table 3: Overview of frequent combinations of explanation aspects, namely, categories, explainability techniques, visualization techniques, and representative papers. For each of the column details refer to section 4.2

Category (#)	Explanation Category	Explanation Approach	Visualization	References
Local Post-Hoc *(8)*	Feature importance	Topic scores, word scores (SHAP), source attribution	Saliency (4), raw declarative representation (1)	(Schaper et al., 2022; Chan et al., 2023; Norkute et al., 2021; Ismail et al., 2023; Vo et al., 2024)
	Provenance	Natural language through knowledge graph	Natural language (1)	(Silva et al., 2019)
	Example driven	Adversarial examples	Natural language (1)	(Ismail et al., 2023)
	Interpretable-by-design	Summarization programs	Raw declarative representation (1)	(Saha et al., 2023)
Local Self-Exp *(9)*	Feature importance	Highlight extraction, interaction matrix, attention scores, injecting human interpretable patterns into attention matrices	Saliency (3), natural language (1)	(Li et al., 2021; Wang et al., 2021; Norkute et al., 2021; Li et al., 2022)
	Surrogate model	Source entailment, keyword extraction, LLM generated rationales, topic modeling	Saliency (2), natural language (1), raw declarative representation (1)	(Kryściński et al., 2020; Reunamo et al., 2022; Jiang et al., 2024; Xie et al., 2024)
	Provenance	Structured opinion graph	Other(1)	(Wang et al., 2020)
Global Post-hoc *(1)*	Feature importance	Mining algorithm to obtain explainable information about sentiment of crowd-sourced reviews	Natural language (1)	(Moody et al., 2022)
Global Self-Exp *(1)*	Feature importance	Interpretable by design	Saliency (1)	(da Silva et al., 2023)

(Ebrahimi et al., 2018) adversarial attacks to generate bounded worst-case perturbations that change the model outcome. Nevertheless, unlike counterfactual examples, adversarial attacks are designed not to obtain meaningful data instances but to obtain imperceptible perturbations, and thus might not give interpretable insights about the model.

Feature importance. Feature importance methods aim to explain the outcome by assigning importance scores to input features, such as lexical features including word/tokens and n-grams, clustering over NN embeddings (Schaper et al., 2022), or manual features obtained from feature engineering. Two popular operations to enable feature importance-based explanations are first-derivative saliency and attention mechanism. Such an approach has been adopted in (Li et al., 2021), where textual features are evaluated and highlighted to explain the generated summary. Soft masking, token-level, and sentence-level extraction help in giving importance scores to the features, thus deciding what features are important to be kept in the summary. Li et al. (2022) employs a human-in-the-loop pipeline, where interpretable patterns identified by humans are injected into the attention matrices of the same or a smaller model. They applied this approach to extractive text summarization, utilizing BERTSum, and reported improvements in the model's interpretability, accuracy, and efficiency.

Surrogate Model. When a surrogate model is used for explainability, the summarization model's outputs are input to the surrogate model, One well-known example is LIME (Ribeiro et al., 2016), which is a model-agnostic method that learns surrogate models using input perturbations. These approaches are model-agnostic and can be used to achieve either local or global explanations. A surrogate model is used in (Reunamo et al., 2022) where they propose an explainable extractor for generating keyword summaries of nursing episodes. To enhance the extraction process, the authors combine a Bidirectional LSTM-based model for text classification with LIME. The LSTM model classifies nursing episodes into different subjects. LIME

is then utilized to explain the classification model's results by identifying the most important words highlighted by the model. These keywords are subsequently extracted and used as the basis for summarization, as they are considered the most central words in each paragraph.

Kryściński et al. (2020) make the important observation that ensuring each summary sentence is entailed by a source sentence helps establish the factual accuracy of the summary, and they train a surrogate model to perform the entailment.

Provenance-based. Provenance-based explanations attempt to illustrate the model's prediction process, where the final prediction is the result of a series of reasoning steps, e.g. Silva et al. (2019) develop a text entailment method in which a natural language explanation is generated along with the model output based on lexical knowledge graph. Wang et al. (2020) presents an interactive review summarization system that provides both a graph-structured summary of the different opinions mentioned in the reviews and a textual summary of the reviews. The system provides the provenance of the opinions presented in the summary by tracing back the original reviews from which opinions were extracted. As an example of an inherently-explainable (self-explaining) summarization model, Saha et al. (2023) propose to generate summaries based on *summarization programs*, binary trees that show how each sentence in the summary was created by referring back to the input sentences.

Insight: Referring to RQ2 from our initial research questions, in Table 3, the feature importance technique is the most extensively used explainability technique (with 8 out of 17 papers). It is well-known that features and their attributions (i.e., quantified importance for the model output) belong to the most reliable explanation aspects for understanding the predictions of black-box models. Other techniques like provenance-based, example-driven, and surrogate models account for 2, 1, and 4 papers respectively.

4.2.3 Visualizations of Explanations (RQ3)

Communicating the explanations visually to the user is a critical part of XAI, since often the users inspecting the explanations are not expected to be ML experts. Generally the data format returned by the explanation method constrains the kinds of visualizations that can be done. Here we give an overview of the common visualizations used across the papers we reviewed.

Saliency maps, in which different parts of the input are highlighted in different intensities corresponding to numerical quantities assigned to them, be it feature importance scores or attention weights, are frequently used for those methods of explanations. Compared to bar charts, saliency maps can be easier to read by embedding the information directly into the input text. Table 3 shows that as feature importance methods and attention scores are frequently used for explanations, saliency maps are the most widely used visualization method.

Raw declarative representations directly visualize the explanation in a data format specific to the method, such as a graph of topics (Wang et al., 2020) or a binary tree showing the relationship between input and summary sentences (Saha et al., 2023).

Natural language explanations that might be generated by another language model or extracted from the input sentence (e.g. keywords) are naturally visualized as text, such as in (Moody et al., 2022).

Other visualization methods. Beyond the above categories of visualization methods, other methods include scoring or inferring the similarity between the generated summary and the input text, as depicted in Fig 1a in the Appendix. Wang et al. (2020) employs a multi-view interactive visualization approach to represent the review summary. Their structured summary utilizes directed edges between nodes, color-coded nodes indicating aspect categories, and font size variations reflecting opinion frequency. The opinions reflected in the generated summary are also color-coded.

Insights: what makes an explanation and its visualization helpful is highly problem-specific and evaluating an explanation's quality is a non-trivial task (Nauta et al., 2023). Since feature importance methods are the most commonly used kind of explanations among the papers we surveyed (Table 3), saliency maps are most frequently used for visualization. While such maps can effectively display keywords or important sentences, they give little insight into the summarization process or the structure between the input/summary sentences. More expressive formats such as graphs (Saha et al., 2023) can be used along with appropriate explanation methods to derive richer insights from the summaries.

4.2.4 Evaluation of Explanations (RQ3)

This section presents how explanations are evaluated in the works we reviewed; we base our categorization on (Danilevsky et al., 2020):

No or informal examination: most reviewed studies don't evaluate the explanations or only provide an informal examination. In some papers, the quality of explanations is assessed based on their impact on summarization task performance, measured through human evaluation (Wang et al., 2021) or metrics such as the ROUGE score and BERTScore (Jiang et al., 2024; Li et al., 2021). This trend is primarily seen in papers where the explanation approach falls into the self-explainable category.

Human evaluation: only two out of 17 studies employ human-based evaluation, involving two and three experts evaluating the explanations of summaries in (Norkute et al., 2021) and (Saha et al., 2023), respectively. This is unsurprising, given the high cost associated with human-based evaluation. In this category, Saha et al. (2023) evaluate the model's immutability, including how well humans can generalize to the model's reasoning patterns with new, unseen inputs based on the provided explanations.

Comparison to ground truth: ground truth evaluation involves comparing the generated explanations with human-annotated textual explanations (Wiegreffe and Marasovic, 2021), considered ground truth for evaluating explanations. This lack of ground-truth evaluation relates to our earlier point in 4.1.2, highlighting the lack of explainable datasets for ATS, where we only encountered one paper. We use this section to reiterate the need to extend the work on constructing datasets with human-annotated explanations for ATS.

Insights: evaluating XAI methods and explanations remains an open challenge in the research field. The lack of evaluation of XAI methods applied to ATS can be attributed to the fact that existing XAI evaluation frameworks primarily focus on computer vision (Hedström et al., 2023; Arras et al., 2022; Kokhlikyan et al., 2020). Those that do support textual use cases mainly focus on classification tasks (Attanasio et al., 2023). However, this is concerning, given research showing that some XAI methods can be unfaithful (Slack et al., 2020; Turpin et al., 2023; Kozik et al., 2024). Therefore, evaluating quality metrics for explanations, such as fidelity, is crucial, especially in high-stakes environments like the ATS of health or legal documents. This aligns with previous calls by the XAI community (Longo et al., 2024; Freiesleben and König, 2023) and should prompt further research on developing evaluation frameworks for XAI methods in NLP, extending current frameworks to tasks like ATS, creating explainable datasets, and facilitating human evaluation studies for explainable NLP.

4.2.5 Conclusions on the Practical Usefulness of Explainability Approaches (RQ4)

Referring to our initial research questions, particularly RQ4, it is evident from our survey that explainability techniques are gaining traction in the field of text summarization. The common use of post-hoc methods (9 out of 19) highlights the community's interest in methods that provide insights after the model predictions to understand and verify model behavior. In this direction, future work on interpreting transformer-based summarization models decisions can include leveraging mechanistic interpretability approaches that focus on reverse engineering a model's decisions and decomposing them into understandable pieces (Templeton et al., 2024; Wang et al., 2023a)

On the other hand, the frequent use of ante-hoc methods (10 out of 19) also indicates the interest in integrating inherent interpretation within the models. This aligns with the increasing focus on developing analysis methods tailored to transformer-based model architectures (Mohebbi et al., 2023a,b)

Moreover, feature importance techniques are most utilized, highlighted in 11 of the 17 surveyed papers. This method is especially valued for its ability to quantify the importance of features in the decisions made by black-box models. Such feature-based approaches are prevalent in text summarization, vision-related, and tabular methods (Borisov et al., 2022), indicating their general reliability and effectiveness in making AI systems more interpretable.

For effective visualization, XAI techniques for text summarization should prioritize simplicity, clarity, and alignment with human intuition. Interactive tools, heatmaps, and consistent visual styles enhance understanding and allow users to explore how different inputs influence model predictions. Scalable visualizations incorporating annotations and clear documentation are crucial for handling complex datasets and ensuring that explanations remain accessible to all users, regardless of their

technical background.

The existing gap in evaluating explanations for ATS can hinder the practical usability of explainability models, especially when summarization is employed in high-stakes environments. As pointed out in 4.2.4, more efforts are necessary to bridge this gap.

Overall, the practical usefulness of explainability approaches in text summarization is increasingly recognized which is essential for building trust and transparency. However, further research is needed to develop comprehensive evaluation frameworks and specialized datasets for explainable text summarization.

4.3 Summarization for Explainability (RQ5)

In this section, we highlight some previous work on how summarization and summaries contribute to explainability.

One way explainability benefits from summaries is by using summaries and summarization in *constructing explainable NLP datasets.* Explainable NLP datasets contain human-annotated textual or human-written justification for the correct label. These datasets exist for various NLP tasks like sentiment classification, claim verification, and question answering. Wiegreffe and Marasovic (2021) reviews and classifies explainable NLP datasets into three categories by explanation type: structured, highlights, and free-text. One example of a dataset that utilizes summaries to construct a free-from explainable dataset for claim verification is LIAR-PLUS (Alhindi et al., 2018), where it contains web-scraped human-written fact-checking summaries that are used as explanations.

Another application direction is using *summarization approaches in the process of generating explanations*; this is primarily seen in fact-checking related work. Atanasova et al. (2020) uses LIAR-PLUS and employs an extractive summarization-based approach to generate veracity explanations where LIAR-PLUS is used as a dataset. Their approach involves training DistilBERT-based models to optimize the extraction of top k sentences similar to the gold justification, where the ROUGE-2 F1 score measures similarity. More recently, Russo et al. (2023) integrates summarization in a claim-driven framework to generate justifications by employing various summarization approaches. They experiment with both extractive and abstractive text summarization methods. Initially, several extractive techniques are applied, followed by a combination of these techniques with an abstractive summarization step performed by different pre-trained language models. This combination achieves the best results when training data is available, highlighting the effectiveness of combining both extractive and abstractive methods compared to using each separately for this task. However, such an approach was still limited to LMs hallucinations.

In the same application direction, Hongwimol et al. (2021) presents a knowledge-graph-based scientific literature discovery platform that provides users with explanations on why certain papers are selected. For each search query and corresponding result, an explanation is attached, detailing the reasons for selecting a particular paper. These explanations are provided in the form of a generated text summary, which utilizes a T5 model to summarize the filtered abstract of the paper based on the user's query. Bacco et al. (2021) employs summarization as a tool to explain the classification outcomes of a hierarchical transformer architecture-based sentiment analysis system for movie reviews. They use transformer-based models for extractive summarization where the most important sentences for the sentiment decision, ranked by attention weights, are used as a basis for the summary.

Text summarization has shown the potential to enhance the interpretability of large language models by facilitating the detection of hallucinations. Identifying when a model has produced a hallucinated output can simplify subsequent explanations of the model's behavior. Vakharia et al. (2024) demonstrate that better summarization ability can also help overcome hallucinations, which is a significant drawback of LLMs, making them harder to trust and, therefore, interpret. Through a dataset of conversations along with their human- and machine-generated summaries and a fine-grained labeling of the hallucinations present, they show that teaching the same seq2seq model to both generate summaries and denote hallucinations (by appending two different heads to the same encoder-decoder model) leads both to better summaries and more accurate detection of hallucinations. While the approach in (Vakharia et al., 2024) can be extended to text-generation tasks beyond summarization, it highlights the synergistic relationship a model's performance has with its interpretability and reliability.

Insights: Summarization has shown its potential in constructing explainable datasets, generating explanations for classification use cases, and im-

proving the interpretability and reliability of LLMs. This highlights the advantages and opportunities for further research that leverages summarization to enhance the interpretability of generative models and other NLP systems.

5 Related Surveys

Two of the earlier baseline surveys in XAI are presented in (Adadi and Berrada, 2018; Guidotti et al., 2018). Adadi and Berrada (2018) serves as a reference for terminologies and approaches regarding XAI and (Guidotti et al., 2018) classifies XAI techniques and provides a comprehensive background regarding the main concepts, motivations, and implications of enabling explainability in intelligent systems. Explainable NLP surveys include (Danilevsky et al., 2020; Zini and Awad, 2022; Luo et al., 2024). Danilevsky et al. (2020) review XAI techniques in NLP with a focus on explaining model's decision for several NLP tasks. Later, Zini and Awad (2022) extends such review by highlighting the explainability methods on the input and processing levels. More recently, Luo et al. (2024) reviews and categorizes the explainability methods specific only for providing local explanations. Focusing on LLMs, (Zhao et al., 2024) overviews and classifies the different approaches for explaining LLMs based on the training paradigms.

6 Conclusion

This paper presents a dual-perspective review of the intersection between XAI and ATS. First, we review the current state of applying XAI to ATS. Second, we highlight the application of summarization in enhancing the interpretability of black-box models. Given our focus on ATS as a use case, this work aims to promote the practical usability of XAI in ATS and other generation tasks in NLP systems. We present this survey as a resource for researchers and practitioners interested in designing, using, or enhancing the explainability of ATS systems. We hope this survey also paves the way for further research into utilizing summarization to improve the interpretability of NLP-based systems.

Future work: To address the urgent need to bridge the gap in ground truth evaluation for explainability methods applied to ATS, future work could focus on designing explainable datasets for text summarization. Motivated by suggestions from (Longo et al., 2024), this could involve augmenting human annotations and rationales with synthetic data to comprehensively evaluate XAI methods for ATS.

7 Limitations

The results, insights, and trends in this paper are primarily based on the reviewed literature at the intersection of XAI and ATS. However, we don't claim to cover *all* the related literature. Our findings may be limited by the scope of the retrieved literature.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions. This research has been supported by the German Federal Ministry of Education and Research (BMBF) grant 01IS23069 Software Campus 3.0 (TU München).

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Leila Arras, Ahmed Osman, and Wojciech Samek. 2022. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. ferret: a framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Luca Bacco, Andrea Cimino, Felice Dell'Orletta, and Mario Merone. 2021. Extractive summarization for explainable sentiment analysis using transformers. In *Sixth International Workshop on eXplainable SENTIment Mining and EmotioN deTection*.

Anna Bodonhelyi, Efe Bozkir, Shuo Yang, Enkelejda Kasneci, and Gjergji Kasneci. 2024. User intent recognition and satisfaction with large language

models: A user study with chatgpt. *arXiv preprint arXiv:2402.02136*.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Hou Pong Chan, Qi Zeng, and Heng Ji. 2023. Interpretable automatic fine-grained inconsistency detection in text summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6433–6444, Toronto, Canada. Association for Computational Linguistics.

Vinícius da Silva, João Paulo Papa, and Kelton Augusto Pontara da Costa. 2023. Extractive text summarization using generalized additive models with interactions for sentence selection. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 737–745. INSTICC, SciTePress.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 229–239.

Timo Freiesleben and Gunnar König. 2023. Dear xai community, we need to talk! In *Explainable Artificial Intelligence*, pages 48–65, Cham. Springer Nature Switzerland.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Trevor Hastie and Robert Tibshirani. 1985. Generalized additive models; some applications. In *Generalized Linear Models*, pages 66–81, New York, NY. Springer US.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. 2023. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11.

Pollawat Hongwimol, Peeranuth Kehasukcharoen, Pasit Laohawarutchai, Piyawat Lertvittayakumjorn, Aik Beng Ng, Zhangsheng Lai, Timothy Liu, and Peerapon Vateekul. 2021. Esra: Explainable scientific research assistant. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 114–121.

Qusai Ismail, Kefah Alissa, and Rehab M. Duwairi. 2023. Arabic News Summarization based on T5 Transformer Approach. In *2023 14th International Conference on Information and Communication Systems (ICICS)*, pages 1–7.

Pengcheng Jiang, Cao Xiao, Zifeng Wang, Parminder Bhatia, Jimeng Sun, and Jiawei Han. 2024. TriSum: Learning Summarization Ability from Large Language Models with Structured Rationale. (arXiv:2403.10351).

Hyun Kim, Minsoo Cho, and Seung-Hoon Na. 2023. ExplainMeetSum: A dataset for explainable meeting summarization aligned with human intent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13079–13098, Toronto, Canada. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

Barbara Ann Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Rafał Kozik, Massimo Ficco, Aleksandra Pawlicka, Marek Pawlicki, Francesco Palmieri, and Michał Choraś. 2024. When explainability turns into a threat - using xai to fool a fake news detection method. *Computers & Security*, 137:103599.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haoran Li, Arash Einolghozati, Srinivasan Iyer, Bhargavi Paranjape, Yashar Mehdad, Sonal Gupta, and Marjan Ghazvininejad. 2021. EASE: Extractive-abstractive summarization end-to-end using the information bottleneck principle. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 85–95, Online and in Dominican Republic. Association for Computational Linguistics.

Raymond Li, Wen Xiao, Linzi Xing, Lanjun Wang, Gabriel Murray, and Giuseppe Carenini. 2022. Human guided exploitation of interpretable attention patterns in summarization and topic segmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10189–10204, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zihao Li. 2023. The dark side of chatgpt: legal and ethical challenges from stochastic parrots and hallucination. *arXiv preprint arXiv:2304.14347*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2024. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 623–631, New York, NY, USA. Association for Computing Machinery.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. 2024. Local interpretations for explainable natural language processing: A survey. *ACM Comput. Surv.*

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*.

Goutam Majumder, Vikrant Rajput, Partha Pakray, Sivaji Bandyopadhyay, and Benoit Favre. 2022. Text summary evaluation based on interpretable semantic textual similarity. *Multimedia Tools and Applications*, pages 1–26.

Michail Mamalakis, Héloïse de Vareilles, Graham Murray, Pietro Lio, and John Suckling. 2024. The explanation necessity for healthcare ai. *arXiv preprint arXiv:2406.00216*.

Lisa Messeri and MJ Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, and Afra Alishahi. 2023a. Homophone disambiguation reveals patterns of context mixing in speech transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8249–8260, Singapore. Association for Computational Linguistics.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023b. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.

Aaron Moody, Chenyi Hu, Huixin Zhan, Makenzie Spurling, and Victor S Sheng. 2022. Towards explainable summary of crowdsourced reviews through text mining. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 528–541. Springer.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.

Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, and Sally Gao. 2021. Towards explainable ai: Assessing the usefulness and impact of added explainability features in legal document summarization. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Akseli Reunamo, Laura-Maria Peltonen, Reetta Mustonen, Minttu Saari, Tapio Salakoski, Sanna Salanterä, and Hans Moen. 2022. *Text Classification Model Explainability for Keyword Extraction – Towards Keyword-Based Summarization of Nursing Care Episodes*, volume 290.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Swarnadeep Saha, Shiyue Zhang, Peter Hase, and Mohit Bansal. 2023. Summarization programs: Interpretable abstractive summarization with neural modular trees. In *The Eleventh International Conference on Learning Representations*.

Ben Schaper, Christopher Lohse, Marcell Streile, Andrea Giovannini, and Richard Osuala. 2022. Towards interpretable summary evaluation via allocation of contextual embeddings to reference text topics. *ArXiv*, abs/2210.14174.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Vinícius Silva, João Papa, and Kelton Costa. 2022. Extractive text summarization using generalized additive models with interactions for sentence selection. *arXiv preprint arXiv:2212.10707*.

Vivian S. Silva, André Freitas, and Siegfried Handschuh. 2019. Exploring knowledge graphs in an interpretable composite approach for text entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7023–7030.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 180–186, New York, NY, USA. Association for Computing Machinery.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Priyesh Vakharia, Devavrat Joshi, Meenal Chavan, Dhananjay Sonawane, Bhrigu Garg, and Parsa Mazaheri. 2024. Don't Believe Everything You Read: Enhancing Summarization Interpretability through Automatic Identification of Hallucinations in Large Language Models. (arXiv:2312.14346).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. *Advances in neural information processing systems*, 30.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. (arXiv:1710.10903).

Song-Nguyen Vo, Tien-Thinh Vo, and Bac Le. 2024. Interpretable extractive text summarization with meta-learning and bi-lstm: A study of meta learning and explainability techniques. *Expert Systems with Applications*, 245:123045.

Haonan Wang, Yang Gao, Yu Bai, Mirella Lapata, and Heyan Huang. 2021. Exploring explainable selection to control abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13933–13941.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023a. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Qiang Wang, Ling Lu, and Aijuan Wang. 2023b. Ltog: An abstractive long-input summarization method based on local-to-global mapping. *Available at SSRN 4538534*.

Xiaolan Wang, Yoshihiko Suhara, Natalie Nuno, Yuliang Li, Jinfeng Li, Nofar Carmeli, Stefanos Angelidis, Eser Kandogann, and Wang-Chiew Tan. 2020. Extremereader: An interactive explorer for customizable and explainable review summarization. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 176–180, New York, NY, USA. Association for Computing Machinery.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Qianqian Xie, Prayag Tiwari, and Sophia Ananiadou. 2024. Knowledge-enhanced graph topic transformer for explainable biomedical text summarization. *IEEE Journal of Biomedical and Health Informatics*, 28(4):1836–1847.

Zebin Yang, Aijun Zhang, and Agus Sudjianto. 2021. Gami-net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, 120:108192.

Huixin Zhan, Kun Zhang, Chenyi Hu, and Victor S. Sheng. 2022. Hgats: hierarchical graph attention networks for multiple comments integration. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, page 159–163, New York, NY, USA. Association for Computing Machinery.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*

Julia El Zini and Mariette Awad. 2022. On the explainability of natural language processing deep models. *ACM Computing Surveys*, 55(5):1–31.

A Review Methodology

For this review, we employed a systematic approach by following the methodology defined by Kitchenham and Charters (2007) with the research questions as:

- **RQ1** What are the popular models, datasets, and evaluation metrics used in existing research on explainable text summarization?
- **RQ2:** What XAI techniques are employed for text summarization in the existing research studies?
- **RQ3:** How are such explanations visualized and evaluated?
- **RQ4:** Can we derive practical conclusions on the usefulness of Explainability techniques for text summarization?
- **RQ5:** How can text summarization methods be utilized by Explainable AI to provide explanations?

To restrict the research scope to the focus of this paper, we then defined a set of related keywords to search popular databases for relevant papers. We applied the following search string to the title, abstract, and keywords: *("explainable" OR "interpretable" OR "explainability" OR "interpretability") AND ("text summarization")*

We queried popular databases databases for relevant papers: ACL anthology, ACM digital library, IEEE Xplore, and Google Scholar.

After obtaining the initial set of papers by applying the search strings, we filtered down the papers based on inclusion and exclusion criteria. Papers were screened for the inclusion criteria: (1) written in English, (2) accessible on the web, (3) papers with a clear focus on text summarization and explainability (4) peer-reviewed papers. We excluded the papers that didn't satisfy all the aforementioned criteria, except very recent pre-prints that satisfied the first three criteria.

After filtering down the papers, we divided the papers into two categories. Papers in first category represent the *Explainability for Text Summarization* direction in which explainability techniques have been applied to text summarization. The second category represents the *Summarization for Explainability* direction and consists of papers where text summarization is used to provide explanations independent of the NLP task under consideration.

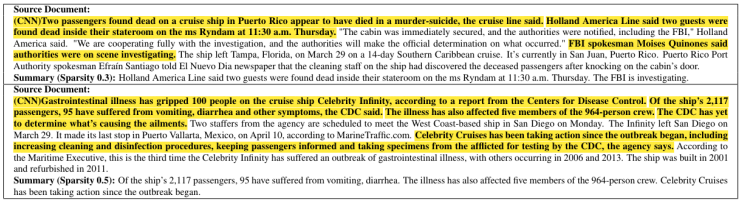
B Additional Figures and Tables

The appendix contains definitions of evaluation metrics for text summarization methods (Table 4), example visualizations of explanations from the reviewed papers (Figure 1), and a list of text summarization datasets used (Table 5).

Table 4: Popular metrics for evaluating text summarization.

Metric	Description
ROUGE Score (Lin, 2004)	N-gram overlap between generated and reference summaries.
BLEU Score (Papineni et al., 2002)	Measure co-occurrence of n-grams in the generated/reference summaries.
METEOR (Lavie and Agarwal, 2007)	Aligns words between the generated/reference summaries for a similarity score.
CIDEr (Lavie and Agarwal, 2007)	Weighting common n-grams based on their rarity in the reference texts.
BERTScore (Zhang et al., 2020b)	Similarity between generated/referenece summaries through BERT embeddings.

(a) Similarity scoring between the summary and input text (Majumder et al., 2022)



(b) Saliency highlighting (Li et al., 2021)

Figure 1: Some examples of visualization techniques of explanations observed in the surveyed papers.

Table 5: Overview of major datasets for text summarization used in the reviewed papers. Publicly available datasets can be accessed by clicking on the dataset's name.

Dataset	Domain	Description	Public
YELP	Business	Reviews and ratings for businesses on Yelp.	✓
CNN/ DailyMail	Journalism	News articles and short summaries.	✓
XSUM	Journalism	News articles and short *abstractive* summaries.	✓
PubMed	Medical	Biomedical and life sciences research articles.	✓
FEVER	General	Fact-checking dataset with claims extracted from Wikipedia.	✓
MNLI	General	Sentence pairs with textual entailment annotations.	✓
Amazon reviews	E-commerce	Customer reviews and ratings on Amazon.	✓
MultiSum	General	Human-validated summaries for texts and videos.	✓
arxiv	Academic	Papers from arXiv.	✓
Aggrefact-Unified	Research	Factuality error annotations separated based on the summary model.	✓
TAC	Academic	Datasets used for various shared tasks including text summarization.	✓
Fake News Corpus	Journalism	News articles known to contain false information.	✓
CORD-19	Academic	Full-text articles on COVID-19 and other coronaviruses.	✓
Nursing Entries	Medical	Nursing entries obtained from a Finnish university hospital.	✗
ClinicalTrials	Medical	Custom-made documents describing the proposal for testing the effectiveness and the safety of a new treatment,	✗
BBC news summary	Multidomain	Documents consisting of news articles and corresponding reference	✓