ISSN: 2178-7727

TRANSFORMER-BASED ABSTRACTIVE TEXT SUMMARIZATION: A COMPREHENSIVE REVIEW OF MODELS, CHALLENGES, AND EVALUATION

Spriha Sinha
Department of Computer Science and Engineering
Government Engineering College, Ambikapur, Digma, Ambikapur
Chhattisgarh Swami Vivekanand Technical University, Bhilai (C.G.), India

Mrs. Pooja Patre
Assistant Professor, Department of Computer Science and Engineering

Abstract

The exponential growth of digital textual information across news portals, academic repositories, legal corpora, and social media platforms has intensified the need for efficient automated text summarization systems. Transformer-based architectures have emerged as the dominant paradigm in abstractive text summarization, offering superior capabilities in modeling long-range linguistic dependencies and generating fluent, coherent summaries. This review paper provides a comprehensive examination of the development, current state, and open challenges of transformer-based abstractive text summarization. We systematically analyze foundational encoder-decoder architectures, pre-trained language models such as BERT, GPT-2, BART, T5, and PEGASUS, and survey critical research contributions addressing hallucination mitigation, factual consistency, long-document processing, and evaluation methodology. We further review prominent benchmark datasets including CNN/Daily Mail, XSum, MultiNews, and ArXiv, and discuss evaluation metrics spanning ROUGE, BERTScore, BARTScore, and NLI-based faithfulness measures. Our analysis identifies persistent challenges in scalable inference, domain adaptation, multilingual summarization, and reliable automatic evaluation, and outlines promising directions for future research in this rapidly evolving field.

Keywords: Abstractive summarization, Transformer, BERT, BART, T5, PEGASUS, ROUGE, hallucination, factual consistency, pre-trained language models, natural language processing, deep learning.

1. Introduction

The unprecedented proliferation of digital content in the contemporary information era has rendered manual reading and comprehension of large text corpora increasingly impractical. From scientific literature and legal documents to social media streams and financial reports, the volume of unstructured textual data generated globally continues to grow at a rate far exceeding human capacity to process it effectively [1]. Automatic text summarization — the task of producing a shorter, coherent, and informationally faithful representation of a source document — has therefore emerged as a critical research domain within Natural Language Processing (NLP) [2].

Text summarization approaches are broadly categorized into extractive and abstractive methods. Extractive summarization selects and assembles salient sentences directly from the source text, while abstractive summarization generates novel sentences that may not appear verbatim in the original document, analogous to how a human expert would write a summary [3]. Abstractive methods are inherently more challenging but produce more natural and concise outputs, making them the focal point of contemporary research.

The development of the Transformer architecture by Vaswani et al. [4] in 2017 marked a watershed moment in NLP. By replacing recurrence with self-attention mechanisms, Transformers enabled models to capture global contextual relationships across entire documents with superior computational efficiency. The subsequent introduction of large-scale pre-trained models — BERT [5], GPT-2 [6], XLNet [7], BART [8], T5 [9], and PEGASUS [10] — leveraging transfer learning on massive corpora further accelerated advances in abstractive summarization.

Despite remarkable achievements, transformer-based summarization models continue to face significant open challenges. These include the generation of factually inconsistent or hallucinated content, the computational intractability of processing very long documents due to quadratic attention complexity, the difficulty of reliable automatic evaluation, limited cross-domain generalizability, and inadequate support for low-resource and multilingual settings [11]. Addressing these challenges constitutes an active and growing body of research.

This review paper provides a structured and comprehensive examination of transformer-based abstractive text summarization. We analyze the evolution from early statistical and neural methods to state-of-the-art pre-trained architectures, survey key research contributions and methodological innovations, discuss benchmark datasets and evaluation frameworks, and delineate open problems and future research directions. Our aim is to serve as a reference for researchers and practitioners entering this field as well as those seeking to situate new contributions within the broader research landscape.

2. Evolution of Text Summarization

2.1 Early Statistical and Rule-Based Methods

The earliest automatic summarization systems, pioneered by Luhn [12] in 1958, employed term frequency statistics to identify salient sentences. Edmundson [13] subsequently introduced additional features such as sentence position, cue phrases, and document structure to improve selection heuristics. These extractive methods were computationally inexpensive but fundamentally limited in their capacity to capture semantic relationships, frequently yielding redundant or incoherent summaries.

Graph-based methods represented a significant advancement in extractive summarization. TextRank [14], inspired by the PageRank algorithm, modeled inter-sentence similarity as a graph and ranked sentences by importance through iterative computation. LexRank [15] extended this paradigm using cosine similarity between TF-IDF sentence vectors. While effective for well-structured texts, graph-based methods remained constrained to extraction and could not produce abstractive paraphrases.

2.2 Machine Learning Approaches

The introduction of supervised machine learning enabled more principled sentence ranking based on annotated training data. Feature-engineered classifiers learned to predict sentence salience using syntactic, lexical, and positional features [16]. Unsupervised clustering methods grouped semantically similar sentences and selected representative cluster centroids. While performance improved over purely rule-based methods, the reliance on hand-crafted features limited generalization across domains and languages [17].

2.3 Neural Sequence-to-Sequence Models

The emergence of deep learning brought a paradigm shift with the introduction of sequence-to-sequence (Seq2Seq) encoder-decoder architectures [18]. Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs) [19], and Gated Recurrent Units (GRUs) enabled the modeling of sequential dependencies in text. Rush et al. [20] demonstrated the first neural abstractive summarization system using attention-based Seq2Seq models for headline generation.

Bahdanau et al. [21] introduced the soft attention mechanism, allowing decoders to dynamically focus on relevant encoder states during generation. See et al. [22] proposed the Pointer-Generator Network, which combined copying from the source document with abstractive generation, significantly reducing hallucination in early neural summarization systems. Despite these advances, RNN-based models suffered from sequential computation bottlenecks and difficulty in modeling long-range dependencies beyond a few hundred tokens.

3. Transformer Architecture for Summarization

3.1 Self-Attention Mechanism

The Transformer architecture introduced by Vaswani et al. [4] fundamentally reconceptualized sequence modeling by replacing recurrence with multi-head self-attention. Self-attention computes pairwise relationships between all token positions simultaneously, enabling the model to capture both local syntactic dependencies and long-range semantic associations within a single computational layer. Multi-head attention extends this by projecting queries, keys, and values into multiple representation subspaces, enriching contextual understanding.

Positional encoding is added to token embeddings to inject sequential order information, which is otherwise absent in the attention-only architecture. The Transformer's feed-forward layers, residual connections, and layer normalization collectively enable stable training of very deep models. The key advantage over recurrent architectures is full parallelizability: all attention computations within a layer can be executed simultaneously, dramatically reducing training time on modern GPU hardware [23].

3.2 Pre-trained Language Models

BERT (Bidirectional Encoder Representations from Transformers) [5] introduced the masked language modeling pre-training objective, enabling bidirectional contextual representations. While BERT's encoder-only architecture is not directly applicable to generation, it influenced the development of encoder representations used in hybrid summarization systems.

Liu and Lapata [24] adapted BERT for extractive summarization by fine-tuning on sentence-level classification tasks.

GPT-2 [6] demonstrated the powerful text generation capabilities of large-scale autoregressive language models trained with next-token prediction. BART [8] combined a bidirectional encoder with an autoregressive decoder pre-trained with a denoising objective, making it particularly well-suited for sequence-to-sequence tasks including abstractive summarization. T5 (Text-to-Text Transfer Transformer) [9] unified all NLP tasks under a text-to-text framework, treating summarization as generating a target text from a source prefix.

PEGASUS [10] introduced the Gap Sentence Generation (GSG) pre-training objective specifically designed for summarization, masking and generating entire salient sentences from documents. This task-aligned pre-training strategy yielded state-of-the-art performance on multiple summarization benchmarks with fewer fine-tuning examples, demonstrating the value of domain-adapted pre-training for downstream task performance.

3.3 Encoder-Decoder Architectures

For abstractive summarization, encoder-decoder Transformer architectures have proven most effective. The encoder processes the input document to produce dense contextual representations, while the decoder generates the summary autoregressively conditioned on encoder outputs through cross-attention. Fine-tuning pre-trained encoder-decoder models on summarization datasets enables transfer of general linguistic knowledge to the specific generation task, reducing data requirements and improving fluency [8, 9].

4. Literature Review

4.1 Hierarchical and Structural Models

Akiyama, Tamura, and Ninomiya [25] proposed Hie-BART, a hierarchical extension of BART specifically designed to address the limitations of flat Transformer architectures in long-document summarization. The model encodes individual sentences before aggregating them into document-level representations, enabling superior discourse modeling and content selection. Experimental results demonstrated improved ROUGE scores over standard BART on long-document benchmarks, confirming that explicit structural modeling enhances summarization coherence.

Koh et al. [26] conducted an empirical survey of long-document summarization methods, taxonomizing approaches into sparse attention mechanisms, hierarchical encoding models, and chunking strategies. Their analysis revealed that standard Transformer attention complexity scales quadratically with input length, fundamentally constraining applicability to long documents. Sparse attention variants such as Longformer [27] and BigBird [28] were evaluated as scalable alternatives, demonstrating that selective attention patterns can maintain competitive performance at reduced computational cost.

4.2 Training Objectives and Optimization

Liu et al. [29] introduced BRIO (Bringing Order to Abstractive Summarization), a contrastive learning framework that addresses the limitations of maximum likelihood estimation (MLE) training. Standard MLE treats reference summaries deterministically and fails to account

for the relative quality of different candidate summaries. BRIO redistributes probability mass across candidates according to their quality scores, training the model to rank better summaries higher during inference. Experiments on CNN/Daily Mail and XSum achieved state-of-the-art results without architectural modifications.

Zhang et al. [30] proposed PEGASUS-X, extending PEGASUS with staggered local-global attention for summarizing long documents. The model interleaves local sliding window attention with global attention tokens to efficiently capture both local coherence and document-level context. Experimental evaluations confirmed competitive performance on long-document benchmarks while maintaining computational tractability.

4.3 Hallucination and Factual Consistency

Hallucination — the generation of content not supported by or contradicting the source document — is among the most critical challenges in abstractive summarization. Maynez et al. [11] provided a foundational empirical characterization of hallucination in summarization, distinguishing intrinsic hallucinations (contradictions of source content) from extrinsic hallucinations (unsupported additions). Their analysis revealed that even state-of-the-art models generate substantial hallucinated content, with different architectures exhibiting distinct error patterns.

Laban et al. [31] developed SummaC, a framework for automatic factual consistency evaluation based on Natural Language Inference (NLI) models. SummaC decomposes summaries into granular text units and checks their consistency against the source document using pre-trained NLI classifiers. Benchmarking studies demonstrated significantly higher correlation with human faithfulness judgments than ROUGE-based metrics, establishing NLI-based evaluation as a more reliable proxy for factual accuracy.

Wan et al. [32] proposed faithfulness-aware decoding strategies that integrate consistency scoring into the beam search process without requiring model retraining. Candidate summaries are re-ranked or penalized during decoding based on entailment scores computed against the source. This inference-time approach significantly reduces factual errors while maintaining high ROUGE performance, offering a practical solution applicable to existing trained models.

Ma et al. [33] introduced BUMP, a meta-evaluation benchmark specifically designed to assess the sensitivity of faithfulness metrics in detecting subtle hallucinations. BUMP constructs minimal pairs of summaries that differ only in factual correctness while maintaining high lexical similarity. Testing established metrics on BUMP revealed that most commonly used measures fail to detect subtle factual errors, identifying a critical gap in current evaluation methodology.

4.4 Evaluation Metrics and Benchmarks

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [34] remains the dominant automatic evaluation metric in summarization research, measuring n-gram and longest common subsequence overlap between generated and reference summaries. However, extensive empirical work has demonstrated that ROUGE correlates poorly with human judgments of factual accuracy, coherence, and overall quality, particularly for abstractive systems that may use different vocabulary to express the same information [35].

Yuan, Neubig, and Liu [36] proposed BARTScore, a generation-probability-based evaluation metric that estimates the likelihood of a candidate summary being generated by a pre-trained BART model conditioned on the source document. This approach jointly captures fluency, relevance, and semantic alignment, demonstrating superior correlation with human judgments compared to ROUGE across multiple summarization tasks. BARTScore has since been adopted as a complement to lexical overlap metrics in rigorous evaluation frameworks.

Li et al. [37] introduced HaluEval, a large-scale hallucination evaluation benchmark providing fine-grained classification of hallucination types across NLP tasks including summarization. HaluEval revealed widespread hallucination in state-of-the-art language models, underscoring the gap between automatic metric scores and factual reliability. Tam et al. [38] conducted empirical evaluations of factual consistency in news summarization, finding significant discrepancies between ROUGE scores and human-assessed factual correctness.

4.5 Entity and Knowledge-Aware Summarization

Zhou et al. [39] developed an entity-aware multi-document summarization framework that explicitly models named entities within the encoder-decoder architecture. By embedding entity representations into cross-document attention, the model maintains entity consistency across source documents, reducing redundancy and factual errors in consolidated summaries. Experimental results on standard multi-document benchmarks demonstrated improved entity coherence and ROUGE performance over baseline models.

Lyu et al. [40] proposed a fact-aware abstractive summarization framework that constructs entity-relation graphs from source documents and integrates them into a pointer-generator network. The hybrid architecture selectively copies factual entities and relations while generating abstractive text, achieving a balance between faithfulness and fluency. Human evaluation confirmed reduced hallucination rates, highlighting the potential of structured knowledge integration for improving factual consistency.

Li et al. [41] introduced boundary-aware summarization with entity-augmented attention, capturing both sentence-level discourse boundaries and entity relationships within a unified attention mechanism. The dual-focus architecture improved factual correctness and structural coherence across automatic and human evaluation measures, with demonstrated applicability to low-resource settings.

4.6 Long-Document and Memory-Efficient Models

Moro et al. [42] proposed a memory-augmented Transformer architecture for long-document summarization in low-resource regimes. An external memory module stores salient representations from document segments, enabling sequential processing of long inputs while maintaining access to global context. The approach achieved competitive summarization performance with substantially reduced memory requirements, extending the practical applicability of Transformer-based summarization to resource-constrained environments.

The Longformer [27] and BigBird [28] architectures addressed the quadratic attention bottleneck through sparse attention patterns combining local sliding window attention with global attention tokens. These models demonstrated effective long-document processing on tasks

including summarization of scientific articles, legal texts, and books, with linear complexity in sequence length enabling processing of tens of thousands of tokens.

4.7 Clinical and Domain-Specific Summarization

Adams, Zucker, and Elhadad [43] conducted a meta-evaluation of faithfulness metrics in the domain of clinical hospital-course summarization, comparing automatic measures against expert human judgments. Their findings revealed that faithfulness metrics validated on news summarization performed poorly in clinical settings, failing to detect subtle medical factual errors. This work highlights the critical importance of domain-specific evaluation and the risks of deploying general-purpose summarization frameworks in high-stakes medical applications without rigorous domain-adapted validation.

5. Datasets and Benchmarks

The CNN/Daily Mail dataset [44] has been the most widely used benchmark for single-document news summarization, containing approximately 300,000 article-highlight pairs. Its moderately extractive nature makes it suitable for evaluating both extractive and abstractive models. XSum [45], in contrast, consists of highly abstractive one-sentence summaries of BBC news articles, challenging models to genuinely compress and paraphrase rather than extract.

For multi-document summarization, the Multi-News [46] dataset provides articles from multiple news sources with human-written summaries. Scientific document summarization has been evaluated on arXiv and PubMed datasets [47], which contain long-form research articles with structured abstracts, presenting significant challenges for models constrained by maximum input sequence lengths.

The BookSum [48] dataset enables evaluation of very long narrative summarization spanning entire book chapters, pushing the boundaries of what Transformer-based models can process. WikiHow [49] provides procedural summarization from instructional articles, testing generalization to non-news domains. Evaluation on diverse datasets remains essential for assessing the true generalizability of summarization models.

6. Challenges and Open Problems

6.1 Computational Scalability

The standard Transformer self-attention mechanism has $O(n^2)$ computational and memory complexity with respect to input sequence length, fundamentally limiting direct processing of long documents. While sparse attention mechanisms provide partial remedies, they introduce approximation trade-offs that may compromise summary quality on documents requiring global coherence across distant sections. Efficient Transformer variants remain an active area of research without definitive solutions [26].

6.2 Hallucination and Factual Reliability

Despite extensive research, hallucination remains a pervasive problem in abstractive summarization systems. Models trained with MLE objectives can generate fluent but factually incorrect summaries, with error rates that remain unacceptably high for deployment in consequential domains such as healthcare, law, and finance [11]. Mitigation strategies including

faithfulness-constrained training, NLI-guided decoding, and structured knowledge integration have shown promise but have not eliminated the problem.

6.3 Evaluation Reliability

The limitations of ROUGE as the primary evaluation metric are well-documented [35]. ROUGE does not measure factual accuracy, semantic coherence, or readability, and correlates poorly with human judgments for abstractive systems. While BERTScore, BARTScore, and NLI-based faithfulness metrics provide complementary perspectives, no single metric captures all dimensions of summary quality. The development of comprehensive, reliable, and computationally tractable evaluation frameworks remains an open challenge [33].

6.4 Domain Adaptation and Multilingual Summarization

Pre-trained models trained predominantly on English news corpora exhibit degraded performance when applied to specialized domains or other languages. Domain-specific summarization in medicine, law, and science requires both domain knowledge and appropriate evaluation standards. Multilingual and cross-lingual summarization presents additional challenges related to training data availability and cross-lingual transfer of summarization capabilities [50].

7. Future Directions

Future research in transformer-based abstractive summarization will likely advance along several interconnected dimensions. First, the development of architectures that combine the efficiency of sparse attention with the quality of full attention — potentially through hierarchical or adaptive attention patterns — will be essential for scaling to longer documents. Second, integrating explicit factuality constraints into training objectives and decoding algorithms represents a critical pathway to improving reliability in high-stakes applications.

Third, the development of robust, domain-adaptive evaluation frameworks that go beyond lexical overlap to assess semantic fidelity, factual accuracy, and readability will be foundational for meaningful research progress. Fourth, advances in multilingual and cross-lingual summarization, leveraging multilingual pre-trained models, will expand the applicability of these systems globally. Fifth, multimodal summarization — integrating textual, visual, and structured data — presents exciting opportunities for richer and more complete information synthesis.

Sixth, the alignment of summarization systems with human preferences through reinforcement learning from human feedback (RLHF) and other preference optimization techniques offers a promising direction for improving the perceived quality and utility of generated summaries. Finally, ensuring the ethical deployment of summarization systems — including transparency, accountability for factual errors, and prevention of misuse for misinformation generation — will be an increasingly important consideration as these systems are integrated into consequential information pipelines.

8. Conclusion

This review has traced the evolution of automatic text summarization from early rule-based and statistical methods through neural sequence-to-sequence models to the current dominant paradigm of transformer-based pre-trained language models. The introduction of architectures

such as BART, T5, and PEGASUS, combined with large-scale pre-training and fine-tuning strategies, has enabled abstractive summarization systems of unprecedented quality and fluency.

However, significant challenges remain. Hallucination and factual inconsistency undermine the trustworthiness of generated summaries in high-stakes applications. Computational constraints limit the processing of long documents. Evaluation metrics fail to capture the full spectrum of summary quality dimensions. Domain adaptation and multilingual generalization remain unsolved. Addressing these challenges through interdisciplinary research combining advances in model architecture, training methodology, evaluation science, and responsible AI will define the trajectory of this field.

The trajectory of progress in transformer-based abstractive summarization is promising. As pre-trained models grow more capable and efficient, as evaluation methodologies mature, and as domain-specific solutions emerge, automated summarization systems will increasingly serve as reliable tools for information access, knowledge synthesis, and decision support across domains.

References

[1] Mani, I. (2001). Automatic summarization. John Benjamins Publishing.

[2] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). Text summarization techniques: A brief survey. International Journal of Advanced Computer Science and Applications, 8(10), 397–405.

[3] Nenkova, A., and McKeown, K. (2011). Automatic summarization. Foundations and Trends in Information Retrieval, 5(2–3), 103–233.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[5] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186.

[6] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8).

[7] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, 32.

[8] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of ACL 2020, 7871–7880.

[9] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

[10] Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. Proceedings of ICML 2020, 11328–11339.

[11] Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. Proceedings of ACL 2020, 1906–1919.

[12] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159–165.

[13] Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM, 16(2), 264–285.

[14] Mihalcea, R., and Tarau, P. (2004). TextRank: Bringing order into texts. Proceedings of EMNLP 2004, 404–411.

[15] Erkan, G., and Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22, 457–479.

[16] Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. Proceedings of SIGIR 1995, 68–73.

[17] Barzilay, R., and Elhadad, M. (1997). Using lexical chains for text summarization. Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 10–17.

[18] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27.

[19] Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

[20] Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. Proceedings of EMNLP 2015, 379–389.

[21] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. Proceedings of ICLR 2015.

[22] See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. Proceedings of ACL 2017, 1073–1083.

[23] Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2022). Efficient transformers: A survey. ACM Computing Surveys, 55(6), 1–28.

[24] Liu, Y., and Lapata, M. (2019). Text summarization with pretrained encoders. Proceedings of EMNLP 2019, 3730–3740.

[25] Akiyama, K., Tamura, A., and Ninomiya, T. (2021). Hie-BART: Document summarization with hierarchical BART. Proceedings of NAACL-SRW 2021.

[26] Koh, H. Y., Ju, J., Liu, M., and Pan, S. (2022). An empirical survey on long document summarization: Datasets, models and metrics. ACM Computing Surveys, 55(8).

[27] Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

[28] Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. (2020). Big Bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33.

[29] Liu, Y., Liu, P., Radev, D., and Neubig, G. (2022). BRIO: Bringing order to abstractive summarization. Proceedings of ACL 2022, 2890–2903.

[30] Zhang, J., Fei, H., and Liu, P. J. (2022). PEGASUS-X: Extending PEGASUS for long input summarization. arXiv preprint arXiv:2208.04347.

[31] Laban, P., Schnabel, T., Bennett, P. N., and Hearst, M. A. (2022). SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. Transactions of the Association for Computational Linguistics, 10, 163–177.

[32] Wan, D., Liu, M., McKeown, K., Dreyer, M., and Bansal, M. (2023). Faithfulness-aware decoding strategies for abstractive summarization. Proceedings of EACL 2023.

[33] Ma, L., Cao, S., Logan IV, R. L., Lu, D., Ran, S., Zhang, K., Tetreault, J., and Jaimes, A. (2023). BUMP: A benchmark of unfaithful minimal pairs for meta-evaluation of faithfulness metrics. Proceedings of ACL 2023.

[34] Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. Proceedings of the ACL Workshop on Text Summarization Branches Out, 74–81.

[35] Tam, D., Mascarenhas, A., Zhang, S., Kwan, S., Bansal, M., and Raffel, C. (2023). Evaluating the factual consistency of large language models through news summarization. Findings of ACL 2023.

[36] Yuan, W., Neubig, G., and Liu, P. (2021). BARTScore: Evaluating generated text as text generation. Advances in Neural Information Processing Systems, 34.

[37] Li, J., Cheng, X., Zhao, W. X., Nie, J. Y., and Wen, J. R. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. Proceedings of EMNLP 2023.

[38] Tam, D., Mascarenhas, A., Zhang, S., Kwan, S., Bansal, M., and Raffel, C. (2023). Evaluating the factual consistency of large language models through news summarization. Findings of ACL 2023.

[39] Zhou, H., Ren, W., Liu, G., Su, B., and Lu, W. (2021). Entity-aware abstractive multi-document summarization. Findings of ACL-IJCNLP 2021.

[40] Lyu, Y., Lu, Z., Lin, C., and Huang, S. (2022). Faithful abstractive summarization via fact-aware entity-relation pointer generator networks. Proceedings of CIKM 2022.

[41] Li, J., Liu, J., Ma, J., Yang, W., and Huang, D. (2024). Boundary-aware abstractive summarization with entity-augmented attention for enhancing faithfulness. ACM Transactions on Asian and Low-Resource Language Information Processing.

[42] Moro, G., Ragazzi, L., Valgimigli, L., Frisoni, G., Sartori, C., and Marfia, G. (2023). Efficient memory-enhanced transformer for long-document summarization in low-resource regimes. Sensors, 23(7), 3542.

[43] Adams, G., Zucker, J., and Elhadad, N. (2023). A meta-evaluation of faithfulness metrics for long-form hospital-course summarization. Proceedings of ML4H 2023.

[44] Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. Advances in Neural Information Processing Systems, 28.

[45] Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. Proceedings of EMNLP 2018, 1797–1807.

[46] Fabbri, A. R., Li, I., She, T., Li, S., and Radev, D. R. (2019). Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. Proceedings of ACL 2019, 1074–1084.

[47] Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. Proceedings of NAACL 2018, 615–621.

[48] Kryscinski, W., Rajani, N., Agarwal, D., Xiong, C., and Radev, D. (2022). BookSum: A collection of datasets for long-form narrative summarization. Findings of EMNLP 2022.

[49] Koupaee, M., and Wang, W. Y. (2018). WikiHow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305.

[50] Hasan, T., Bhattacharjee, A., Islam, M. S., Samin, K., Li, Y., Kang, Y. B., Rahman, M. S., and Shahriyar, R. (2021). XL-Sum: Large-scale multilingual abstractive summarization for 44 languages. Findings of ACL-IJCNLP 2021, 4693–4703.