The current state of summarization

Fabian Retkowski

1. Introduction

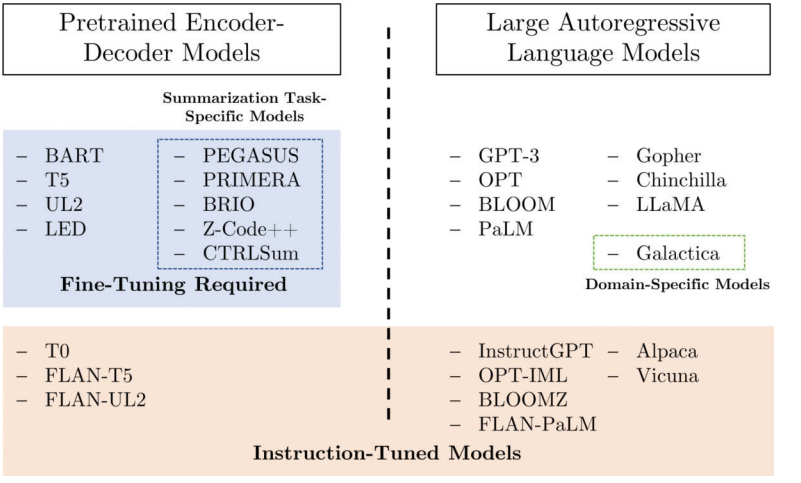
Summarization is the process of extracting the most important information from a text and presenting it in a condensed form. With vast amounts of information produced at an unprecedented rate, organizations and individuals alike face unique challenges, heightening the demand for effective summarization systems. For researchers of many fields, it is challenging to keep up with the latest developments in their field including Artificial Intelligence itself as vicariously indicated by the number of journal publications per year which has almost tripled since 2015 (D. Zhang et al. 2022).

In general, two different forms of summarization are distinguished: extractive and abstractive. In extractive summarization, the system is tasked with selecting passages from the document to be included in the summary. Abstractive summarization, on the other hand, aims to rephrase the most important aspects of a document with a different syntax. As language models are becoming more and more capable, research is increasingly shifting from extractive to abstractive summarization, which is considered more challenging, but also more fluent, diverse, and readable.

This paper covers recent advances in abstractive text summarization, with a focus on pre-trained encoder-decoder models (Section 2), large autoregressive language models (Section 3), and instruction-tuned variants (Section 4). While aiming to be reasonably comprehensive, Figure 1 gives an overview of the covered models. In Section 5, current evaluation protocols are discussed in the context of the paradigm shift towards large language models. At the end of the paper, we discuss limitations, potentials (Section 6), and current commercialization efforts (Section 7).

2. Pre-trained encoder-decoder models

Figure 1: Current summarization systems can be broadly divided into pre-trained encoder-decoder models and large autoregressive language models. In general, instruction-tuned models are most capable when it comes to zero-shot summarization. Other encoder-decoder models usually require fine-tuning, while autoregressive LLMs are less effective without instruction tuning. Illustration courtesy of the author.



Pre-trained encoder-decoder models have gained tremendous popularity in recent years and are now widely established in the field of natural language processing. These models are trained in a self-supervised setting on a large, unlabeled corpus. Notable examples include models such as the denoising autoencoder BART (Lewis et al. 2020) and T5 (Raffel et al. 2020) that is trained on a fill-in-the-blank objective. UL2 (Tay et al. 2022) serves as a more recent example that generalizes and combines several denoising pre-training objectives. By fine-tuning these models on task-specific datasets, they have achieved state-of-the-art results across many tasks including summarization. Some pre-trained models are specifically designed for the task of summarization by choosing a pre-training objective that resembles summarization. For example, in Figure 2, the architecture of PEGASUS (J. Zhang et al. 2020) can be observed, which is trained by removing important sentences from the input document and tasking the model with regenerating them. In a comprehensive

evaluation of 23 models for the summarization task, Fabbri et al. (2021: 400) conclude that PEGASUS, BART, and T5 "consistently performed the best on most dimensions", which involves human evaluations as well as automatic metrics. Recently, a task-specific fine-tuning mechanism called BRIO (Liu et al. 2022) was proposed for summarization. This method introduces a contrastive learning component to prevent assigning the entire distribution mass to the reference summary and instead account for candidate summaries as well. BRIO has been applied to several models, including BART and PEGASUS. Another noteworthy model is Z-Code++ (P. He et al. 2023), as it incorporates an intermediate task-adaptive fine-tuning step using a broad collection of summarization datasets before fine-tuning on a specific summarization task. This method has been shown to be especially effective in low-resource settings.

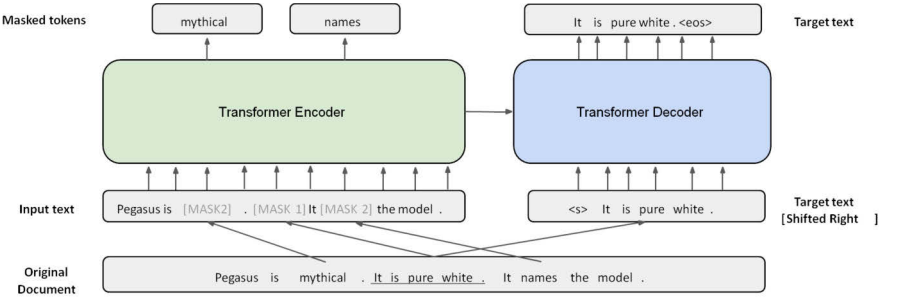
3. Large autoregressive language models

Another significant paradigm shift is the recent emergence of large autoregressive language models (LLMs). These decoder-only models tend to have many more parameters and are trained using the traditional causal language modeling objective of predicting the next token in a sequence. Brown et al. (2020) were the first to demonstrate that this approach, at scale, enables zero-shot prompting to perform a wide variety of downstream tasks. Without any gradient updates, this involves priming the model with a task-specific natural language prompt (e.g., "Question: question Answer:") and then producing an output by sampling from the model. The same paradigm also allows for zero-shot summarization, which can be achieved by appending "TL;DR:" to a prompt, among other options.

The most popular model in this category is GPT-3 (Brown et al. 2020) with its 175B parameters. OPT (S. Zhang et al. 2022) and BLOOM (BigScience Workshop 2022) are two open-source alternatives aimed to replicate the results. Gopher (Rae et al. 2022) and PaLM (Chowdhery et al. 2022) take this approach to the extreme by scaling to even larger model sizes of up to 560B parameters. On the contrary, Chinchilla (Hoffmann et al. 2022) and LLaMA (Touvron et al. 2023) take scaling laws and compute budgets more strictly into consideration and this way achieve training a 70B respectively 65B model while still being able to match or outperform larger models. It is also worth mentioning the Galactica 120B scientific language model (Taylor et al. 2022), which demonstrates the effectiveness of specialized LLMs. It outperforms other LLMs in its spe-

cific domain by using a sophisticated dataset design that incorporates domain-adapted tokenization. It treats citations and modalities such as chemical formulas and protein sequences in a special manner by introducing task-specific tokens for them.

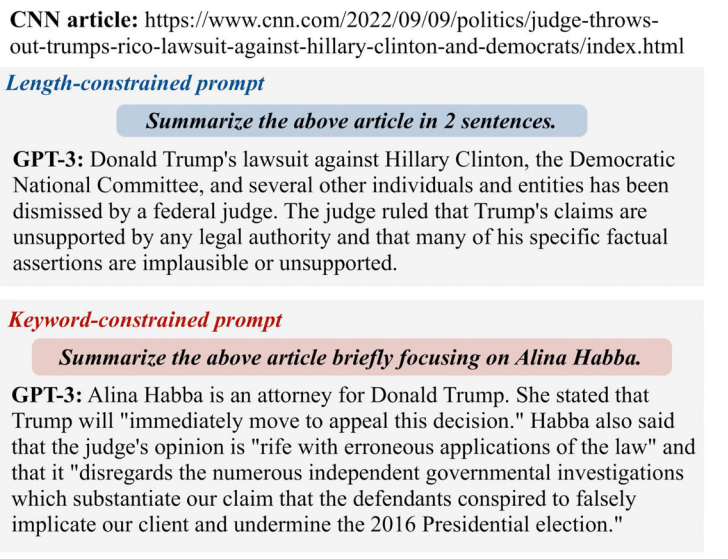
Figure 2: The PEGASUS architecture with its pre-training objectives. The model combines Masked Language Modeling (MLM) as well as Gap Sentences Generation (GSG). As part of GSG, important sentences are masked and used as a target for the decoder. The importance is proximately scored by ROUGE-1 between a sentence and the remaining portions of the document. Taken from J. Zhang et al. 2020.



4. Instruction-tuned models

Instruction tuning refers to the process of fine-tuning a pre-trained model with a diverse range of datasets that are described using natural language task instructions. This step ensures that the training process is more aligned with how the model will be used during inference and has been shown to significantly improve performance on zero-shot tasks. It enables the model to be straightforwardly and more reliably instructed to perform a certain task. For instance, it is now possible to use "Summarize the article: article" as a prompt for the summarization task. More prompt examples are shown in Figure 3.

Figure 3: Exemplary instructions for zero-shot summarization using GPT-3. Notably, the natural language instructions of LLMs enable greater control over tasks, such as length-constrained summarization. Taken from Goyal/Li/Durrett 2022.



To tune models for instructions, the most common approaches are supervised fine-tuning and reinforcement learning from human feedback (RLHF, Christiano et al. 2017). When it comes to pre-trained encoder-decoder models, there are several popular instruction-tuned models available. For instance, T0 (Sanh et al. 2022) and FLAN-T5 (Chung et al. 2022), which are both based on T5, have gained significant traction among practitioners. The same is true for large autoregressive language models of which most have an instruction-tuned equivalent: InstructGPT (Ouyang et al. 2022), OPT-IML (Iyer et al. 2023), BLOOMZ (Muennighoff et al. 2023), FLAN-PaLM (Chung et al. 2022). Taylor et al. (2022) demonstrated with Galactica an alternative approach to enable rudimentary instruction prompting with their prompt pre-training method. This involves adding task prompts to the pre-training, rather than tuning the model after pre-training. A recent trend in the open-source community is to fine-tune LLMs based on conversational and instruction-following data generated by an existing and strong instruction-tuned LLM such as ChatGPT. This has led to the

development of Alpaca and Vicuna, both of which are based on LLaMA (Taori et al. 2023; The Vicuna Team 2023; Y. Wang et al. 2023). The task of summarization is represented in most natural-language-prompted datasets. For example, in the API prompt dataset used by InstructGPT, 4.2% of instructions fall under the 'summarization' use case. Similarly, T0 augments classic summarization datasets like CNN Daily Mail (Nallapati et al. 2016) or SamSum (Gliwa et al. 2019) with instruction templates that can be used to fine-tune the model.

5. Evaluation of large language models

Most commonly, summarization systems are evaluated on automated metrics. ROUGE (Lin 2004) in particular has a long-standing history in the field and measures the lexical overlap between reference summaries and generated summaries. More recent metrics such as BertScore (Zhang et al. 2019) and BARTScore (Yuan/Neubig/Liu 2021), which are better at capturing semantic equivalence, are also becoming increasingly established. However, as large language models become more capable and generalize to a wide range of tasks, they are less frequently or thoroughly evaluated on summarization tasks specifically. Instead, they are evaluated on benchmark suits that focus on question answering and common-sense reasoning, such as SuperGLUE (A. Wang et al. 2019) or MMLU (Hendrycks et al. 2020), that do not explicitly involve summarization. As a result, several research groups have independently investigated the capabilities and limitations of LLMs in summarization more recently (Goyal/Li/Durrett 2022; Bhaskar/Fabbri/Durrett 2023; Liu et al. 2023; Qin et al. 2023; Xiao et al. 2023; Yang et al. 2023; T. Zhang et al. 2023). According to Goyal, Li, and Durrett (2022), summaries generated by instruction-tuned GPT-3 receive lower scores on automatic metrics compared to fine-tuned encoder-decoder models (T0 and BRIO). Despite this, the model outperforms them significantly in human evaluation. The conducted human evaluation by T. Zhang et al. (2023) suggests that they even surpass the reference summaries in quality and are on par with high-quality summaries collected separately for this evaluation. These works cast great doubt on existing evaluation protocols, especially in the context of this paradigm shift. Several of the works describe the low correlation of automatic metrics with human judgment, low reference quality, lacking inter-annotator agreement, and different summarization styles (in length, abstractiveness, formality) as problematic. This is in line with issues raised in previous works such as Fabbri et al. (2021) that point

out the lack of comparability of summarization evaluation protocols – for automated metrics and human evaluation alike. Considering these issues and with summarization systems rivaling human performance, T. Zhang et al. (2023: 10) hypothesize that a limit is reached in evaluating "single-document news summarization", while Yang et al. (2023: 5) call for "rethinking further directions for various text summarization tasks". In fact, the "glass ceiling" phenomenon has been observed more broadly in natural language generation, with even recent automated metrics barely improving correlation with human judgment (Colombo et al. 2022).

6. Limitations and new frontiers

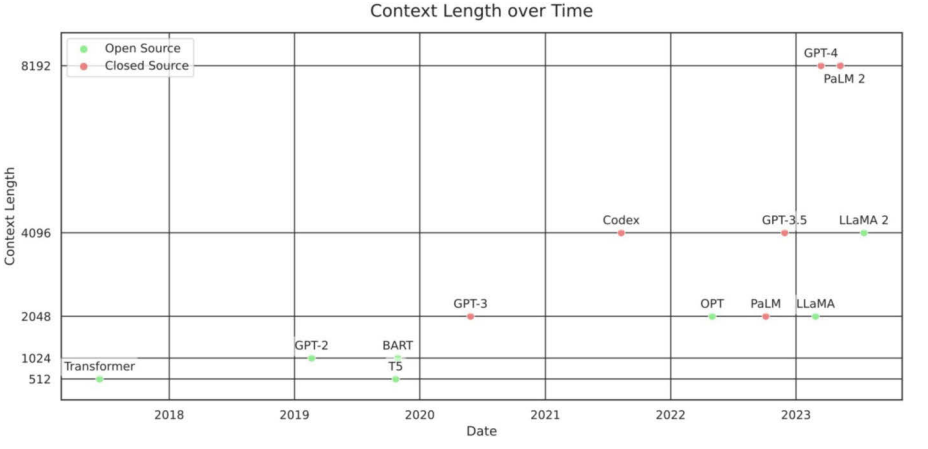
As discussed, there are severe limitations to the current evaluation metrics and protocols, and finding a new standard is an essential area for future research. Liu et al. (2023), for example, suggest using atomic facts to reduce ambiguity in human evaluation, while a recent work in the area of machine translation shows that LLMs themselves make state-of-the-art evaluators offering greater correlation with human judgment than any other automatic metric (Kocmi/Federmann 2023). The latter is also supported by Kadavath et al. (2022), who find that LLMs are capable of self-evaluation. At the same time, LLMs are known to suffer from hallucinations (Ji et al. 2023) and as summarization moves to higher levels of abstractiveness, factuality comes into question. Works like Bhaskar/Fabbri/Durrett (2023) or Goyal/Li/Durrett (2022) show that summarization factuality is still an unsolved issue for LLMs, while others openly discuss how to measure factuality in the first place (Kryscinski et al. 2020; Pagnoni/Balachandran/Tsvetkov 2021).

6.1 Long document summarization

Despite exponential progress (see Figure 4), many current summarization systems are still hindered by the limited context windows of language models which prevent them from processing longer documents that would especially benefit from summarization such as lengthy news articles, scientific papers, podcasts, or books. There are several common strategies to overcome this limitation. One simple method involves truncating the input text (Zhao/Saleh/Liu 2020; A. Wang et al. 2022). For some document types such as news articles, this might serve as a reasonable strategy, as they tend to convey the most salient

information in the beginning. In fact, selecting the first k sentences (Lead-k) is often used as a baseline summary for news summarization systems (See/Liu/Manning 2017; Zhong et al. 2019). In a similar vein, for the summarization of scientific papers, often only the abstract, introduction, and conclusion (AIC) are passed to the summarizer, as previous research found these sections to be the most salient (Sharma/Li/Wang 2019; Cachola et al. 2020). Another approach is to employ an extractive summarizer or retrieval module such as Dense Passage Retriever, Karpukhin et al. (2020), as part of a two-stage system, to select important segments before passing the text to the abstractive summarizer (Liu/Lapata 2019b; Ladhak et al. 2020; A. Wang et al. 2022). There are also transformer architectures that do not suffer from these limitations such as LED (Beltagy/Peters/Cohan 2020) or LongT5 (Guo et al. 2022) which replace $O(n^2)$ attention patterns with more efficient ones. Finally, experiments have been conducted on summarizing chunks of the text in potentially multiple iterations before producing a final, coherent summary (Gidiotis/Tsoumakas 2020; Zhao/Saleh/Liu 2020; Wu et al. 2021; Y. Zhang et al. 2022; Yang et al. 2023).

Figure 4: The context length has been steadily and exponentially increasing in open-source and closed-source language models alike. Not considered are models like LED, which specifically try to maximize the context length at the cost of performance otherwise. Illustration courtesy of the author.



6.2 Multi-document summarization

The process of creating a summary from a collection of documents related to a specific topic is called multi-document summarization (MDS). This presents similar challenges to summarizing a long document, as the problem of limited context length is amplified when multiple documents are involved. Understanding the relationships between the documents is also essential for completing the task effectively. The first strategy for MDS is to simply concatenate all documents into one large text and use techniques designed for single-document summarization. However, this requires the model to process very long sequences. Therefore, a two-stage process similar to that used for long document summarization is commonly employed (Liu et al. 2018; Liu/Lapata 2019a). State-of-the-art approaches also use hierarchical architectures or graph-based methods to capture inter-document relations (Liu/Lapata, 2019a; W. Li et al. 2020; Pasunuru et al. 2021). At the same time, MDS approaches increasingly aim to utilize pre-trained encoder-decoder models such as BART, T5, or PEGASUS (Goodwin/Savery/Demner-Fushman 2020; Pasunuru et al. 2021). One recent and noteworthy model in this category, PRIMERA, is specifically designed for MDS and builds upon the foundations laid by PEGASUS (Xiao et al. 2022). For the GSG objective, PRIMERA chooses sentences that represent clusters of documents. It employs a document concatenation approach and architecturally uses LED to handle long sequences. In this manner, the model is generally applicable, and there are no dependencies on specific datasets. Although there is no scientific evaluation yet, the recent emergence and popularity of practical tools like LangChain and LlamaIndex hint towards the use of LLMs to handle collections of documents. For instance, LlamaIndex enables the storage of documents in an index that is organized like a tree, with each node representing a summary of its child nodes.

6.3 Controllable summarization

Controllable summarization is a multifaceted research question that refers to both the form or style (such as length, formality, or abstractiveness) and the content of a summary. The summary may be conditioned on a specific aspect or entity or, more broadly, on any given keyword or query. In recent years, a wide variety of approaches have been proposed. One of the most comprehensive systems is CTRLSum (J. He et al. 2022), a pre-trained encoder-decoder that generalizes controllability by utilizing keywords and prompts alike. In evaluations,

the authors show the effectiveness of their method for length and entity control, as well as some more specialized tasks (e.g., patent purpose summarization). Recent studies conducted by Goyal/Li/Durrett (2022), Xiao et al. (2023), and Yang et al. (2023) offer initial insights into the potential of instruction-tuned LLMs like GPT-3 and ChatGPT. These systems have shown great promise for diverse summarization tasks based on keywords, aspects, and queries. Figure 3 shows two examples of how zero-shot prompting can enable controllable summarization in such systems. Nevertheless, the potential of LLMs for this task is still largely unexplored. Yang et al. (2023) note that their results can only serve as a lower bound, as the models are naively prompted without any prompt tuning or self-correction. A first glimpse of the potential of a more sophisticated prompting strategy is provided by Xiao et al. (2023) who suggest editing generated summaries with an editor model based on instructions from a separately trained model. In stark contrast, there is also a significant amount of research that focuses on controlling only one aspect of summarization. For example, in length-controllable summarization alone, systems have been proposed that early-stop the decoding process (Kikuchi et al. 2016), select information before passing it to the summarizer (LPAS; Saito et al. 2020), or incorporate length information as part of the input (Kikuchi et al. 2016; Liu/Luo/Zhu 2018). More recently, Liu, Jia, and Zhu (2022) also introduced a length-aware attention mechanism (LAAM).

6.4 Multi-modal summarization

So far, most research attention has been given to text summarization systems. However, there is an abundance of media and content such as podcasts, movies, and meetings that not only involve text but also other modalities including images, videos, and audio. These other modalities potentially contain key information that a pure text summarization system might miss, thus creating a semantic gap. For instance, H. Li et al. (2017) have demonstrated the importance of including audio and video information in the task of summarizing multimedia news, while the work of M. Li et al. (2019) has shown the value of including participants' head orientation and eye gaze when summarizing meetings. One of the key challenges of multi-modal summarization systems is the fusion of different input modalities. Currently, most systems take a late-fusion approach (see Jangra et al. 2023), for example by utilizing a pre-trained encoder. However, recently, a number of promising Transformer-based models have been proposed, which allow the input of diverse modalities

such as Perceiver IO (Jaegle et al. 2021) or GATO (Reed et al. 2022) that have yet to be applied for the summarization task.

7. Commercialization

With language models having surpassed a certain level of performance, the creation and integration of these models into products and tools have become increasingly common, leading to a "gold rush" of NLP startups (Butcher 2022; Toews 2022). For summarization systems in particular, the context lengths of models are of utmost importance and have expanded exponentially in recent years as can be seen in Figure 4, to a level that is practical for more tasks and commercially viable. As such, many summarization systems have become productized and have been made available in consumer-oriented interfaces over the past year. In 2022, Google introduced document summarization in Google Docs (Saleh/Kannan 2022) and conversation summarization in Google Chat (Saleh/Wang 2022), both powered by fine-tuning the PEGASUS model. However, low-quality summaries in the datasets are mentioned as problematic. To tackle this issue, the developers utilize techniques such as dataset distillation, data formatting, and clean-ups, while continuing to collect more training data. Through knowledge distillation, they distill the models into more efficient hybrid architectures of a transformer encoder and a recurrent neural network (RNN) decoder. Separately, an additional model is trained to filter out generated summaries that are of low quality. More recently, Microsoft announced plans to roll out meeting summarization powered by GPT-3.5 in Microsoft Teams in Q2 2023 (Herskowitz 2023), but they have not provided any further technical details. Discord, the community messaging platform, uses "OpenAI technology" for grouping messages into topics for conversation summaries (Midha 2023). Zoom's recent smart recording feature, which includes meeting summarization and smart chaptering, vaguely mentions the use of GPT-3 to "augment" its own models (Parthasarathy 2023). Cohere just launched a dedicated text summarization endpoint (Hillier/Gallé 2023) that largely avoids several problems of LLMs such as the need for prompt engineering and limited context length. In addition, they offer settings to gain more control over the generated summaries: the level of extractiveness, the length, and the format (either fluent text or bullet points). More broadly, access to any standard LLM naturally allows for summarization by specifying the respective prompt. This is true for OpenAI's GPT-3, AI21 Studio, Antrophic's Claude, or

Cohere Generate – to name some that are available via paid APIs and power summarization functionalities in many commercial applications. ChatGPT might be notable, as it also enables a more interactive approach to summarization. Domain-specific summarization tools are another area of interest. For instance, Zoom IQ for Sales (Larkin 2022) aims to provide insights and summaries for sales meetings, while BirchAI, a spinoff from the Allen Institute for Artificial Intelligence, focuses solely on providing customer call summaries for call centers. Meanwhile, beyond big tech and distinguished AI labs, summarization systems are starting to reach many more surfaces such as browsers (Opera; Szyndzielorz 2023), email clients (Shortwave; Wenger 2023) or note-taking apps (Notion; I. Zhao 2023). This trend suggests that summarization is not an application on its own, but a basic feature to be widely implemented on most surfaces and to be widely accessible in the foreseeable future.

8. Conclusion

Text summarization is a rapidly evolving field with two recent paradigm shifts. First, towards finetuning pre-trained encoder-decoder models, and second and even more recently, towards zero-shot prompting of instruction-tuned language models. As a result of these developments, it appears that single-document summarization has reached a tipping point where the focus on improving automated metrics has diminishing returns and might even misdirect the research community. Therefore, we suggest a shift of emphasis towards improving human evaluation protocols and exploring self-evaluation of LLMs. Additionally, more targeted evaluation of certain aspects, such as factuality, should be considered and more broadly the uncovering of capabilities of pre-trained language models and fine-tuned summarization models. However, when contemplating summarization in a wider scope, tasks such as multi-document summarization and multi-modal summarization continue to present significant hurdles. Nonetheless, abstractive text summarization systems for single documents have matured and are rapidly being integrated into consumer products.

List of references

Beltagy, Iz/Peters, Matthew E./Cohan, Arman (2020): Longformer: The Long-Document Transformer, arXiv Preprint (http://arxiv.org/abs/2004.05150).

Bhaskar, Adithya/Fabbri, Alex/Durrett, Greg (2023): "Prompted Opinion Summarization with GPT3.5." In: Findings of the Association for Computational Linguistics (ACL 2023), Toronto, Canada, pp. 9282–9300.

BigScience Workshop (2022): BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, arXiv Preprint (http://arxiv.org/abs/2211.05100).

Brown, Tom/Mann, Benjamin/Ryder, Nick/Subbiah, Melanie/Kaplan, Jared D./Dhariwal, Prafulla/Neelakantan, Arvind/et al. (2020): "Language Models are Few-Shot Learners." In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Vancouver, Canada (https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html).

Butcher, Mike (2022): "Here's why a gold rush of NLP startups is about to arrive.", July 28, 2022 (https://techcrunch.com/2022/07/28/a-gold-rush-of-nlp-startups-is-about-to-arrive-heres-why/).

Cachola, Isabel/Lo, Kyle/Cohan, Arman/Weld, Daniel (2020): "TLDR: Extreme Summarization of Scientific Documents." In: Findings of the Association for Computational Linguistics (EMNLP 2020), online, pp. 4766–4777.

Chowdhery, Aakanksha/Narang, Sharan/Devlin, Jacob/Bosma, Maarten/Mishra, Gaurav/Roberts, Adam/Barham, Paul/et al. (2022): PaLM: Scaling Language Modeling with Pathways, arXiv Preprint (http://arxiv.org/abs/2204.02311).

Christiano, Paul F./Leike, Jan/Brown, Tom/Martic, Miljan/Legg, Shane/Amodei, Dario (2017): "Deep Reinforcement Learning from Human Preferences." In: Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA (https://papers.nips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html).

Chung, Hyung Won/Hou, Le/Longpre, Shayne/Zoph, Barret/Tay, Yi/Fedus, William/Li, Yunxuan/et al. (2022): Scaling Instruction-Finetuned Language Models, arXiv Preprint (http://arxiv.org/ abs/2210.11416).

Colombo, Pierre/Peyrard, Maxime/Noiry, Nathan/West, Robert/Piantanida, Pablo (2022): The Glass Ceiling of Automatic Evaluation in Natural Language Generation, arXiv Preprint (http://arxiv.org/abs/2208.14585).

Fabbri, Alexander R./Kryściński, Wojciech/McCann, Bryan/Xiong, Caiming/Socher, Richard/Radev, Dragomir (2021): "SummEval: Re-evaluating Sum-

marization Evaluation." In: Transactions of the Association for Computational Linguistics 9, pp. 391–409.

Gidiotis, Alexios/Tsoumakas, Grigorios (2020): "A Divide-and-Conquer Approach to the Summarization of Long Documents." In: IEEE/ACM Transactions on Audio, Speech, and Language Processing 28, pp. 3029–3040.

Gliwa, Bogdan/Mochol, Iwona/Biesek, Maciej/Wawer, Aleksander (2019): "SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization." In: Proceedings of the 2nd Workshop on New Frontiers in Summarization, Hong Kong, China, pp. 70–79.

Goodwin, Travis/Savery, Max/Demner-Fushman, Dina (2020): "Flight of the PEGASUS? Comparing Transformers on Few-shot and Zero-shot Multi-document Abstractive Summarization." In: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), pp. 5640–5646.

Goyal, Tanya/Li, Junyi Jessy/Durrett, Greg (2022): News Summarization and Evaluation in the Era of GPT-3, arXiv Preprint (http://arxiv.org/abs/2209.12356).

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

He, Junxian/Kryscinski, Wojciech/McCann, Bryan/Rajani, Nazneen/Xiong, Caiming (2022): "CTRLsum: Towards Generic Controllable Text Summarization." In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, pp. 5879–5915.

He, Pengcheng/Peng, Baolin/Wang, Song/Liu, Yang/Xu, Ruochen/Hassan, Hany/Shi, Yu/et al. (2023): "Z-Code++: A Pretrained Language Model Optimized for Abstractive Summarization." In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, pp. 5095–5112.

Hendrycks, Dan/Burns, Collin/Basart, Steven/Zou, Andy/Mazeika, Mantas/Song, Dawn/Steinhardt, Jacob (2020): "Measuring Massive Multitask Language Understanding." In: 9th International Conference on Learning Representations (ICLR 2021), Virtual Event (https://openreview.net/forum?id=d7KBjmI3GmQ).

Herskowitz, Nicole (2023): "Microsoft Teams Premium: Cut costs and add AI-powered productivity.", February 1, 2023 (https://www.microsoft.com/

en-us/microsoft-365/blog/2023/02/01/microsoft-teams-premium-cut-costs-and-add-ai-powered-productivity/).

Hillier, Sheena/Gallé, Matthias (2023): "Introducing Cohere Summarize Beta: A New Endpoint for Text Summarization.", February 22, 2023 (https://txt.cohere.ai/summarize-beta/).

Hoffmann, Jordan/Borgeaud, Sebastian/Mensch, Arthur/Buchatskaya, Elena/Cai, Trevor/Rutherford, Eliza/de Las Casas, Diego/et al. (2022): Training Compute-Optimal Large Language Models, arXiv Preprint (https://arxiv.org/abs/2203.15556).

Iyer, Srinivasan/Lin, Xi Victoria/Pasunuru, Ramakanth/Mihaylov, Todor/Simig, Daniel/Yu, Ping/Shuster, Kurt/et al. (2023): OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization, arXiv Preprint (http://arxiv.org/abs/2212.12017).

Jaegle, Andrew/Borgeaud, Sebastian/Alayrac, Jean-Baptiste/Doersch, Carl/Ionescu, Catalin/Ding, David/Koppula, Skanda/et al. (2021): "Perceiver IO: A General Architecture for Structured Inputs & Outputs." In: The Tenth International Conference on Learning Representations (ICLR 2022), Virtual Event (https://openreview.net/forum?id=fILj7WpI-g).

Jangra, Anubhav/Mukherjee, Sourajit/Jatowt, Adam/Saha, Sriparna/Hasanuzzaman, Mohammad (2023): A Survey on Multi-modal Summarization, arXiv Preprint (http://arxiv.org/abs/ 2109.05199).

Ji, Ziwei/Lee, Nayeon/Frieske, Rita/Yu, Tiezheng/Su, Dan/Xu, Yan/Ishii, Etsuko/et al. (2023): "Survey of Hallucination in Natural Language Generation." In: ACM Computing Surveys 55/12 (https://doi.org/10.1145/3571730).

Kadavath, Saurav/Conerly, Tom/Askell, Amanda/Henighan, Tom/Drain, Dawn/Perez, Ethan/Schiefer, Nicholas/et al. (2022): Language Models (Mostly) Know What They Know, arXiv Preprint (http://arxiv.org/abs/2207.05221).

Karpukhin, Vladimir/Oguz, Barlas/Min, Sewon/Lewis, Patrick/Wu, Ledell/Edunov, Sergey/Chen, Danqi/Yih, Wen-Tau (2020): "Dense Passage Retrieval for Open-Domain Question Answering." In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, pp. 6769–6781.

Kikuchi, Yuta/Neubig, Graham/Sasano, Ryohei/Takamura, Hiroya/Okumura, Manabu (2016): "Controlling Output Length in Neural Encoder-Decoders." In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, pp. 1328–1338.

Kocmi, Tom/Federmann, Christian (2023): Large Language Models Are State-of-the-Art Evaluators of Translation Quality, arXiv Preprint (http://arxiv.org/abs/2302.14520).

Kryscinski, Wojciech/McCann, Bryan/Xiong, Caiming/Socher, Richard (2020): "Evaluating the Factual Consistency of Abstractive Text Summarization." In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, pp. 9332–9346.

Ladhak, Faisal/Li, Bryan/Al-Onaizan, Yaser/McKeown, Kathleen (2020): "Exploring Content Selection in Summarization of Novel Chapters." In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, pp. 5043–5054.

Larkin, Theresa (2022): "Zoom IQ for Sales: Conversational intelligence for sellers.", April 13, 2022 (https://blog.zoom.us/zoom-iq-for-sales/).

Lewis, Mike/Liu, Yinhan/Goyal, Naman/Ghazvininejad, Marjan/Mohamed, Abdelrahman/Levy, Omer/Stoyanov, Veselin/Zettlemoyer, Luke (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, Online, pp. 7871–7880.

Li, Haoran/Zhu, Junnan/Ma, Cong/Zhang, Jiajun/Zong, Chengqing (2017): "Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video." In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, pp. 1092–1102.

Li, Manling/Zhang, Lingyu/Ji, Heng/Radke, Richard J. (2019): "Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization." In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 2190–2196.

Li, Wei/Xiao, Xinyan/Liu, Jiachen/Wu, Hua/Wang, Haifeng/Du, Junping (2020): "Leveraging Graph to Improve Abstractive Multi-Document Summarization." In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, pp. 6232–6243.

Lin, Chin-Yew (2004): "ROUGE: A Package for Automatic Evaluation of Summaries." In: Text Summarization Branches Out. Proceedings of the ACL-04 Workshop, Barcelona, Spain, pp. 74–81.

Liu, Peter J./Saleh, Mohammad/Pot, Etienne/Goodrich, Ben/Sepassi, Ryan/Kaiser, Lukasz/Shazeer, Noam (2018): "Generating Wikipedia by Summarizing Long Sequences." In: 6th International Conference on Learning Rep-

resentations (ICLR 2018), Vancouver, Canada (https://openreview.net/forum?id=HygovbWC-).

Liu, Yang/Lapata, Mirella (2019a): "Hierarchical Transformers for Multi-Document Summarization." In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 5070–5081.

Liu, Yang/Lapata, Mirella (2019b): "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), Hong Kong, China, pp. 3730–3740.

Liu, Yixin/Fabbri, Alex/Liu, Pengfei/Zhao, Yilun/Nan, Linyong/Han, Ruilin/Han, Simeng/et al. (2023): "Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation." In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, pp. 4140–4170.

Liu, Yixin/Liu, Pengfei/Radev, Dragomir/Neubig, Graham (2022): "BRIO: Bringing Order to Abstractive Summarization." In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, pp. 2890–2903.

Liu, Yizhu/Jia, Qi/Zhu, Kenny (2022): "Length Control in Abstractive Summarization by Pretraining Information Selection." In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, pp. 6885–6895.

Liu, Yizhu/Luo, Zhiyi/Zhu, Kenny (2018): "Controlling Length in Abstractive Summarization Using a Convolutional Neural Network." In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 4110–4119.

Midha, Anjney (2023): "Discord is Your Place for AI with Friends.", March 13, 2023 (https://discord.com/blog/ai-on-discord-your-place-for-ai-with-friends).

Muennighoff, Niklas/Wang, Thomas/Sutawika, Lintang/Roberts, Adam/Biderman, Stella/Le Scao, Teven/Bari, M. Saiful/et al. (2023): "Crosslingual Generalization through Multitask Finetuning." In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, pp. 15991–16111.

Nallapati, Ramesh/Zhou, Bowen/dos santos, Cicero Nogueira/Gulcehre, Caglar/Xiang, Bing (2016): Abstractive Text Summarization Using Se-

quence-to-Sequence RNNs and Beyond, arXiv Preprint (http://arxiv.org/abs/1602.06023).

Ouyang, Long/Wu, Jeffrey/Jiang, Xu/Almeida, Diogo/Wainwright, Carroll/Mishkin, Pamela/Zhang, Chong/et al. (2022): "Training language models to follow instructions with human feedback." In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, USA/Online, pp. 27730–27744.

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

Parthasarathy, Vijay (2023): "Zoom's AI innovations empower people.", February 24, 2023 (https://blog.zoom.us/ai-driven-innovations/).

Pasunuru, Ramakanth/Liu, Mengwen/Bansal, Mohit/Ravi, Sujith/Dreyer, Markus (2021): "Efficiently Summarizing Text and Graph Encodings of Multi-Document Clusters." In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, pp. 4768–4779.

Qin, Chengwei/Zhang, Aston/Zhang, Zhuosheng/Chen, Jiaao/Yasunaga, Michihiro/Yang, Diyi (2023): Is ChatGPT a General-Purpose Natural Language Processing Task Solver?, arXiv Preprint (http://arxiv.org/abs/2302.06476).

Rae, Jack W./Borgeaud, Sebastian/Cai, Trevor/Millican, Katie/Hoffmann, Jordan/Song, Francis/Aslanides, John/et al. (2022): Scaling Language Models: Methods, Analysis & Insights from Training Gopher, arXiv Preprint (http://arxiv.org/abs/2112.11446).

Raffel, Colin/Shazeer, Noam/Roberts, Adam/Lee, Katherine/Narang, Sharan/Matena, Michael/Zhou, Yanqi/Li, Wei/Liu, Peter J. (2020): "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." In: Journal of Machine Learning Research 21/140, pp. 1–67.

Reed, Scott/Zolna, Konrad/Parisotto, Emilio/Colmenarejo, Sergio Gómez/Novikov, Alexander/Barth-Maron, Gabriel/Giménez, Mai/et al. (2022): "A Generalist Agent." In: Transactions on Machine Learning Research (https://openreview.net/forum?id=1ikK0kHjvj).

Saito, Itsumi/Nishida, Kyosuke/Nishida, Kosuke/Otsuka, Atsushi/Asano, Hisako/Tomita, Junji/Shindo, Hiroyuki/Matsumoto, Yuji (2020): Length-

controllable Abstractive Summarization by Guiding with Summary Prototype, arXiv Preprint (http://arxiv.org/abs/2001.07331).

Saleh, Mohammad/Kannan, Anjuli (2022): "Auto-generated Summaries in Google Docs.", March 23, 2022 (https://ai.googleblog.com/2022/03/auto-generated-summaries-in-google-docs.html).

Saleh, Mohammad/Wang, Yinan (2022): "Conversation Summaries in Google Chat.", November 18, 2022 (https://ai.googleblog.com/2022/11/conversation-summaries-in-google-chat.html).

Sanh, Victor/Webson, Albert/Raffel, Colin/Bach, Stephen H./Sutawika, Lintang/Alyafeai, Zaid/Chaffin, Antoine/et al. (2022): Multitask Prompted Training Enables Zero-Shot Task Generalization, arXiv Preprint (http://arxiv.org/abs/2110.08207).

See, Abigail/Liu, Peter J./Manning, Christopher D. (2017): "Get To The Point: Summarization with PointerGenerator Networks." In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, pp. 1073–1083.

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

Szyndzielorz, Julia (2023): "Opera enters the generative AI space with new features in browsers and content apps.", February 10, 2023 (https://blogs.opera.com/news/2023/02/opera-aigc-integration/).

Taori, Rohan/Gulrajani, Ishaan/Zhang, Tianyi/Dubois, Yann/Guestrin, Carlos/Liang, Percy/Hashimoto, Tatsunori B. (2023): "Alpaca: A Strong, Replicable Instruction-Following Model.", March 13, 2023 (https://crfm.stanford.edu/2023/03/13/alpaca.html).

Tay, Yi/Dehghani, Mostafa/Tran, Vinh Q./Garcia, Xavier/Wei, Jason/Wang, Xuezhi/Chung, Hyung Won/et al. (2022): "UL2: Unifying Language Learning Paradigms." In: The Eleventh International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda (https://openreview.net/forum?id=6ruVLB727MC).

Taylor, Ross/Kardas, Marcin/Cucurull, Guillem/Scialom, Thomas/Hartshorn, Anthony/Saravia, Elvis/Poulton, Andrew/Kerkez, Viktor/Stojnic, Robert (2022): Galactica: A Large Language Model for Science, arXiv Preprint (http://arxiv.org/abs/2211.09085).

The Vicuna Team (2023): "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.", March 30, 2023 (https://lmsys.org/blog/2023-03-30-vicuna).

Toews, Rob (2022): "A Wave Of Billion-Dollar Language AI Startups Is Coming.", March 27, 2022 (https://www.forbes.com/sites/robtoews/2022/03/27/a-wave-of-billion-dollar-language-ai-startups-is-coming/).

Touvron, Hugo/Lavril, Thibaut/Izacard, Gautier/Martinet, Xavier/Lachaux, Marie-Anne/Lacroix, Timothee/Rozière, Baptiste/et al. (2023): LLaMA: Open and Efficient Foundation Language Models, arXiv Preprint (https://arxiv.org/abs/2302.13971).

Wang, Alex/Pang, Richard Yuanzhe/Chen, Angelica/Phang, Jason/Bowman, Samuel R. (2022): "SQuALITY: Building a Long-Document Summarization Dataset the Hard Way." In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, pp. 1139–1156.

Wang, Alex/Pruksachatkun, Yada/Nangia, Nikita/Singh, Amanpreet/Michael, Julian/Hill, Felix/Levy, Omer/Bowman, Samuel (2019): "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems." In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, Canada (https://proceedings.neurips.cc/paper_files/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html).

Wang, Yizhong/Kordi, Yeganeh/Mishra, Swaroop/Liu, Alisa/Smith, Noah A./Khashabi, Daniel/Hajishirzi, Hannaneh (2023): "Self-Instruct: Aligning Language Models with Self-Generated Instructions." In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, pp. 13484–13508.

Wenger, Jacob (2023): "AI Email Summaries: Read emails in seconds.", February 27, 2023 (https://www.shortwave.com/blog/ai-email-summaries/).

Wu, Jeff/Ouyang, Long/Ziegler, Daniel M./Stiennon, Nisan/Lowe, Ryan/Leike, Jan/Christiano, Paul (2021): Recursively Summarizing Books with Human Feedback, arXiv Preprint (http://arxiv.org/abs/2109.10862).

Xiao, Wen/Beltagy, Iz/Carenini, Giuseppe/Cohan, Arman (2022): "PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization." In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, pp. 5245–5263.

Xiao, Wen/Xie, Yujia/Carenini, Giuseppe/He, Pengcheng (2023): ChatGPT-steered Editing Instructor for Customization of Abstractive Summarization, arXiv Preprint (http://arxiv .org/abs/2305.02483).

Yang, Xianjun/Li, Yan/Zhang, Xinlu/Chen, Haifeng/Cheng, Wei (2023): Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization, arXiv Preprint (http://arxiv.org/abs/2302.08081).

Yuan, Weizhe/Neubig, Graham/Liu, Pengfei (2021): "BARTScore: Evaluating Generated Text as Text Generation." In: Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Online, pp. 27263–27277.

Zhang, Daniel/Maslej, Nestor/Brynjolfsson, Erik/Etchemendy, John/Lyons, Terah/Manyika, James/Ngo, Helen/et al. (2022): The AI Index 2022 Annual Report, arXiv Preprint (http://arxiv.org/ abs/2205.03468).

Zhang, Jingqing/Zhao, Yao/Saleh, Mohammad/Liu, Peter (2020): "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization." In: Proceedings of the 37th International Conference on Machine Learning, Online, pp. 11328–11339.

Zhang, Susan/Roller, Stephen/Goyal, Naman/Artetxe, Mikel/Chen, Moya/Chen, Shuohui/Dewan, Christopher/et al. (2022): OPT: Open Pre-trained Transformer Language Models, arXiv Preprint (http://arxiv.org/abs/2205.01068).

Zhang, Tianyi/Kishore, Varsha/Wu, Felix/Weinberger, Kilian Q./Artzi, Yoav (2019): BERTScore: Evaluating Text Generation with BERT, OpenReview Preprint (https://openreview .net/forum?id=SkeHuCVFDr).

Zhang, Tianyi/Ladhak, Faisal/Durmus, Esin/Liang, Percy/McKeown, Kathleen/Hashimoto, Tatsunori B. (2023): Benchmarking Large Language Models for News Summarization, arXiv Preprint (http://arxiv.org/abs/2301.13848).

Zhang, Yusen/Ni, Ansong/Mao, Ziming/Wu, Chen Henry/Zhu, Chenguang/Deb, Budhaditya/Awadallah, Ahmed/Radev, Dragomir/Zhang, Rui (2022): "Summ^N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents." In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, pp. 1592–1604.

Zhao, Ivan (2023): "Notion AI is Here, for Everyone.", February 22, 2023 (https://www.notion.so/blog/notion-ai-is-here-for-everyone).

Zhao, Yao/Saleh, Mohammad/Liu, Peter J. (2020): SEAL: Segment-wise Extractive-Abstractive Long-form Text Summarization, arXiv Preprint (http://arxiv.org/abs/2006.10213).

Zhong, Ming/Liu, Pengfei/Wang, Danqing/Qiu, Xipeng/Huang, Xuanjing (2019): "Searching for Effective Neural Extractive Summarization: What Works and What's Next." In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 1049–1058.