

Investigating Fine-tuning of Large Language Models for Text Summarisation

Usama Khaliq¹, Preeti Patel¹,

¹ London Metropolitan University, 166-220 Holloway Road, London
{ usakha123@gmail.com, p.patel@londonmet.ac.uk }

Abstract. This research aim was to investigate whether fine-tuning medium-sized Large Language Models on text summarisation across different fields, from news to scientific papers and governmental reports, would improve the quality of the summaries generated in the given field. The performance was evaluated quantitatively using ROUGE and BLEU scores and qualitatively through human analysis and judgement. However, fine-tuning led to minimal performance improvements compared to the untrained models. The generated summaries were generally accurate; surprisingly, one model introduces false information. In contrast, bigger models, like GPT-4, with its 1.7 trillion parameters, generated near-perfect summaries without having been trained on a specific dataset. The limited performance increase from the fine-tuned models was likely due to small datasets and medium-sized LLMs. Deploying and fine-tuning these models is computationally expensive, requiring careful planning and optimisation. Overall, this project provided insights into transformer fine-tuning, capabilities and limitations that can guide future work to adapt these powerful models for real-world applications.

Keywords: Large Language Models (LLMs), BART, PEGASUS, ProphetNet, Text summarization, ROUGE BLEU

1 Introduction

Summarisation is considered one of the essential skills for a human due to its ability to communicate complex information concisely and understandably without losing the original meaning. It is a skill taught to children since the early years of schooling and is essential in both personal and professional life. In an academic context, summarisation is necessary for students to demonstrate their ability to comprehend and draw conclusions; however, to effectively summarise a text, it is necessary to understand the content on a deeper level, directly developing critical thinking. In the Big Data and information era this skill, is more vital than ever as we are surrounded by data, and summarising it, not necessarily in writing, but also mentally aids comprehension. The rapid evolution of Large Language Models (LLMs), such as

OpenAI's GPT-4 or Google's PaLM, is transforming the way we interact with information. LLMs have applications not only in computer related fields but in a variety of sectors, from healthcare and customer service to marketing and communication. Large Language Models can be compared to a Swiss knife, as they can perform multi-faceted tasks, from generating new text, translating from one language to another, answering questions or summarising lengthy documents. However, each application requires a specific fine-tuning to harness its full potential

2 Related Work

A comprehensive survey of automated text summarization methods [15] concludes that to produce higherquality and robust summaries, intelligent integration with other systems is required. The various types of summaries can be categorised as extractive, abstractive, informative, and indicative [16]. Extractive, as the name suggests, are summaries containing parts of the original document without any modifications, while abstractive summaries include elements not present in the original record. Those parts are obtained through different methods, such as ad aggregation, deletion, or generation. Before the development of large language models (LLMs), the focus was on automating summaries using extraction; compared to abstraction, it was a relatively easy task, as the process would extract the most important parts and merge them. The disadvantage of extractive summaries is that not all parts may be relevant to the summary. Informative summaries provide information, while indicative or critical summaries use the author's viewpoint to write a summary. However, an automated summary will never be purely indicative or informative but somewhere in between. To determine the quality of a summary, a benchmarking framework called SUMMAC (summary content evaluation) [14] was created, it allowed to evaluate summaries to be generated automatically. SUMMAC consists of 4 steps, preparation, choosing a source document, and writing a summary; the last step is summarisation, a summary generated automatically for the selected document. Lastly, the generated summaries are compared using different metrics such as precision, f score or recall. The evaluation results are analysed to determine the performance of each system. Ghadimi and Beigy, introduce SGCSumm [4]. This novel extractive multi-document summarisation technique combines graph theory, a branch of mathematics that studies graphs. These are structures used to model relations between objects, submodularity, the propriety of set functions that provide performance guarantees for the optimisation problem, and deep learning. It has three main components: representation learning, sentence scoring, and summary generation. SGCSumm guarantees a minimum performance bound and nearoptimal solutions by preserving properties like monotonicity and submodularity. Experiments show that its performance can be compared with state-of-the-art summarisers. There are different metrics, such as BLEU and ROUGE, to evaluate the quality of a test generated by LLMs. However, those metrics diverge, in dimensions like coherence, grammar and subjectiveness,

from what a human considers "good" texts. A group of researchers proposed a framework called DRPE [23], which uses roleplayer prompting to assess objectivity by checking grammar, coherence and subjectivity and creating a roleplayer that mimics the perspective of potential readers. The roleplayers need to be made depending on the input text. On one text, multiple roleplayers can work simultaneously. However, unnecessary roleplayers are removed via clustering and remaining to evaluate the text. Experiments on two summarization datasets reveal that the proposed DRPEScore correlates more with human judgment than existing automatic metrics and a basic LLM-based metric. Including both objective and subjective roleplayers yields the best performance. This LLM-based evaluation framework offers a more human-like evaluation ability, particularly for summarization tasks. The performance of a language model depends not only on the input text but also on the model itself; researchers from Stanford and Columbia University [26] have fine-tuned and tested ten different large language models on news summarization tasks. They systematically evaluated how other factors, like tuning instructions, model scale and in-context learning, influence the overall outcome. This led to two key observations. First, instruction tuning, not the scale of the model, is crucial to LLMs' zero-shot summarisation capability. The researchers find that instruction-tuned models, especially GPT-3's Davinci model, perform best despite having fewer parameters than other non-instructiontuned LLMs. This indicates that instruction tuning is more important than the model scale for zero-shot summarisation. Second, existing benchmarks like CNN/DailyMail and XSUM have low-quality reference summaries that underestimate human performance and the potential of finetuned and few-shot LLMs. The researchers recruited freelance writers to create higher-quality summaries for 100 news articles. They find that reference summaries in these benchmarks are judged to be of worse quality than even the best LLM summaries. With the higher-quality freelance writer summaries, the researchers evaluated GPT-3 Davinci against the writers. They find out that while Davinci summaries are much more extractive, paraphrasing less and copying more from the article, crowdsourced annotators rate Davinci and freelance writer summaries as comparable overall quality. However, individual reviewers show varying preferences, indicating subjectivity in human evaluation. Another study [12] compared text generated by large language models, in particular, ChatGPT and a text written by students; the aim was to determine whether ChatGPT was able to generate reflective writing responses and whether those responses could be distinguished from student-written reflections. The outcome was that ChatGPT could generate high-quality reflective responses that scored higher than studentwritten reflections based on six assessment criteria: returning to experience, attending to feelings, association, integration, validation and appropriation. This suggests ChatGPT's capability for reflective writing tasks. Human evaluators were not able to differentiate between ChatGPT-generated text and student-written text. However, BERT, a classifier trained with domain-specific data was able to distinguish ChatGPT-generated reflections from student-written ones with high accuracy, outperforming human evaluators and general-domain AI text detectors. This shows the potential of classifiers trained in a particular area. While ChatGPT's capabilities for reflective writing are promising, educators must be cautious of its

potential negative impacts and ensure it is appropriately used to develop students' reflective and critical thinking skills.

3 Preliminaries

This section describes the preparations before starting the fine-tuning process, including researching the models, datasets, metrics, and environment requirements.

3.1 Models

The requirements for choosing the models were that they must be between 400 and 600 million number parameters. The reason for this choice, was due to the high computational resources necessary to finetune a model. The chosen models were Facebook's BART(Bidirectional and Auto-Regressive Transformers) [11], Google's PEGASUS(Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence) [25], and Microsoft's ProphetNet [19].

3.2 Datasets

The chosen datasets were also medium-sized, between 10k and 100k instances. The reason for this was the same as the models. All the datasets were chosen from the Huggingface library; the chosen models are: • pszemraj/scientific_lay_summarisation – plos – norm, the PLOS dataset is one of the two datasets introduced in the paper "Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature" [5] is a collection of biomedical research paper with their relative summaries, the articles in the dataset were sourced from the Public Library of Science (PLOS), an open-source library. The summaries in the dataset are written manually by the authors, and in the lay version, the summaries are written using non-technical summaries. • multi_news dataset was introduced in the paper "Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model" [3]; it was the first large-scale multidocument summarization. It contains 56,216 news with their relative summaries written by editors; the articles cover various topics and sources. • ccdv/govreport – summarization was introduced in the paper "Efficient Attentions for Long Document Summarization" [9]; the dataset contains 19.5k U.S. government reports with expertwritten summaries, each report is 9.4k words long.

3.3 Metrics and Evaluation

The evaluation process is twofold. Initially, the performance of the models will be quantified using ROUGE and BLEU scores, as this will give an objective measure of how good the model is at generating summaries. BLEU stands for Bilingual Evaluation Understudy; it is used as a quantitative measure of the quality of generated summaries. Originally developed to be used for translation between two languages[24], it works by comparing n-grams in the generated summaries to the

n-grams in a human-created reference summary; an ngram is a sequence of n items from a sample of text; for example, 1-gram of "the Earth is round" is "the" "Earth" "is" "round", while 2-gram is "the Earth" "is round". The score oscillated between 0 to 1, where 1 is a perfect match; it gives an objective measure without considering any semantic meaning. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation is used to evaluate the quality of summaries by comparing them to reference summaries[24]; there are different types of ROUGE scores:

- ROUGE-N measures the overlap in n-gram;
- ROUGE-1 is the overlap of single words, while
- ROUGE-2 is the overlap of 2-word sequences.
- ROUGE-L measures the longest common subsequence between the generated summary and the reference.
- ROUGE-S measures the skip-bigram, any pair of words in their sentence order, allowing for arbitrary gaps.

Afterwards, a qualitative analysis using human judgment will be conducted. This step involves a human reading the original texts and the generated summaries and assessing their quality. This gives a subjective measure of the model's performance and highlights issues the objective evaluation metrics may have missed. Aspects of this analysis include:

- Accuracy: the summaries should explain the main idea, and details from the original text, without introducing incorrect information.
- Completeness: the summaries should include all key points from the original text.
- Relevance: only the key points should be summarised.
- Coherence: the summaries should be logically valid and easy to understand.
- Conciseness: the summaries should be as short as possible without redundancy.
- Clarity: the summaries should be clear and easy to understand.

3.4 Environment

In machine learning, new and more advanced models are created daily using PyTorch and transformers libraries. Python is the preferred programming language for developing and fine-tuning these models due to its easy-to-understand syntax and the availability of powerful machine-learning libraries like TensorFlow, PyTorch and sci-kit-learn. In the end, it is up to the developer to decide. In this case, a docker container with Python as the programming language and PyTorch and HuggingFace transformers libraries were used. But the most crucial factor was having hardware with enough power. Machine Learning tasks heavily rely on GPUs' processing power because they can perform multiple thousands of calculations simultaneously, unlike CPUs which are limited by the number of cores. More than consumer-grade hardware is often required to fine-tune large language models adequately.

4 Experimentation

The exploration of the effects of fine-tuning Large Language Models (LLMs) on domain-specific datasets is based on fine-tuning the three LLMs, BERT, PEGASUS and ProphetNet, on three different datasets, scientific_lay_summarisation-PLoS-norm, multi_news and gov report-summarisation. The goal was to determine if the three models' performance would increase when fine-tuned on a dataset from a specific field. Before starting the fine-tuning process, the three chosen models were evaluated using the same criteria to which the fine-tuned models were subject. This was done to compare if there were any improvements between the base model and the fine-tuned models and to create a base benchmark to which to compare the overall result. After the fine-tuned models were tested on unseen text, text coming from different fields, news, governmental reports and scientific literature, the chosen texts were:

- Kerch bridge is hated symbol of Russian occupation [1]
- South Korea floods: Tunnel horror brings home climate fears [13]
- How climate change worsens heatwaves, droughts, wildfires and floods [20]
- Higher Education Policy Statement and Reform [7]
- HMRC's annual report and accounts 2022 to 2023: performance overview [8]
- Civil Aviation Authority review: report [6]
- The ocean's color is changing as a consequence of climate change [2]
- ChatGPT designs a robot [21]
- Deep-sleep brain waves predict blood sugar control [18]

One of the most challenging parts of fine-tuning a model is finding the right hyperparameters through trial and error. This is a time-consuming task, which initially was not successful. Therefore, the decision was to use the already tested hyperparameters [10]. Some of the crucial parameters are learning rate, weight decay, number of epochs, number of data loaders and precision. The learning rate controls the step size taken at each iteration while moving toward the global minima; this parameter is important because if it is too low, the model risks converging slowly and could underfit, while on the other hand, if it is too high could cause the overshoot, therefore will be unable to converge leading to poor generalisation. One strategy to prevent overfitting is to add a weight decay function, which penalises the loss function.

5 Findings

Interestingly the untrained BART Large Language model, as shown in Tables 1 and 2, has a higher BLEU and ROUGE score than the fine-tuned model. In the tables below, it is possible to see all metrics in play. In the testing splits of the three datasets, the ROUGE scores of the untrained models are around 10% higher than those of the fine-tuned models, while the BLEU scores show a significant difference in favour of the untrained BART model with scores ranging from 50% up to 169%. To understand the meaning of the values, we refer to the paper "Adaptations of ROUGE and BLEU to Better Evaluate Machine Reading Comprehension Task" [24], which explains that the closer the scores are to 1, the better the generated summary is, 1 means that there is a perfect match between the generated summary and the human reference. As expected, the PEGASUS fine-tuned models (Table 4) outperformed the untrained ones (Table 3), even though the difference between the model types regarding ROUGE scores is minimal. However, BLEU scores are almost double in the fine-tuned variants. This could indicate that the models generate many correct phrases matching the reference summaries. However, they are missing important information in the references, resulting in a lower ROUGE score. The Fine-tune ProphetNet (Table 6) outputted a result similar to PEGASUS, outperforming the untrained models (Table 5).

Table 1. BART scores on test split on Untrained Models

Dataset	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Multi News	0.3501	0.1179	0.1953	0.0424
Gov Report	0.2475	0.0903	0.1421	0.0091
Med Paper	0.3734	0.1179	0.2023	0.0472

Table 2. BART scores on test split on Fine-Tuned Models

Dataset	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Multi News	0.3509	0.1306	0.1978	0.0385
Gov Report	0.2434	0.0986	0.1504	0.0052
Med Paper	0.3882	0.1300	0.2195	0.0482

Table 3. PEGASUS scores on test split on Untrained Model

Dataset	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Multi News	0.3185	0.1054	0.1869	0.0214
Gov Report	0.2006	0.0744	0.1251	0.0012
Med Paper	0.3435	0.1127	0.1970	0.0325

Table 4. PEGASUS scores on test split on Fine-Tuned Model

Dataset	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Multi News	0.3375	0.1272	0.1946	0.0355
Gov Report	0.2529	0.1000	0.1556	0.0068
Med Paper	0.3950	0.1349	0.2240	0.0509

Table 5. ProphetNet scores on test split on Untrained Model

Dataset	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Multi News	0.3185	0.1054	0.1869	0.0214
Gov Report	0.2006	0.0744	0.1251	0.0012
Med Paper	0.3435	0.1127	0.1970	0.0325

Table 6. ProphetNet scores on test split on Fine-Tuned Model

Dataset	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Multi News	0.3375	0.1272	0.1946	0.0355
Gov Report	0.2529	0.1000	0.1556	0.0068
Med Paper	0.3950	0.1349	0.2240	0.0509

Training loss and validation loss in the fine-tuning of the three models for the Multi news dataset showed a decreasing trend across the 3 epochs, however, the difference between them is negligible, which could mean that the models might not have generalized enough. This applies also to the metrics, which have not improved significantly, as observed in Tables 7, 8 and 9.

Table 7. Average scores on Validation split of the three Fine-tuned models on Multi News

Epoch	Training Loss	Validation Loss	ROUGE-1	ROUGE-2
1	2.551833	2.387316	0.2949	0.1040
2	2.468766	N/A	0.2889	0.1021
3	2.3977	N/A	0.2893	0.1020

ROUGE-L	ROUGE-L SUM	BLEU	Gen Len
0.1796	0.1796	0.2942	103.6531
0.1769	0.1772	0.2783	99.5758
0.1769	0.1772	0.2786	99.3754

Table 8. Average scores on Validation split of the three Fine-tuned models on Gov Report

Epoch	Training Loss	Validation Loss	ROUGE-1	ROUGE-2
1	2.222000	2.044120	0.3253	0.1413
2	2.068066	2.327420	0.3364	0.1503

3	1.994733	1.984313	0.3438	0.1538
---	----------	----------	--------	--------

ROUGE-L	ROUGE-L SUM	BLEU	Gen Len
0.2203	0.2204	0.3071	129.6502
0.2292	0.2293	0.3273	118.2924
0.2338	0.2339	0.3396	111.2887

Table 9. Average scores on Validation split of the three Fine-tuned models on Med Paper

Epoch	Training Loss	Validation Loss	ROUGE-1	ROUGE-2
1	2.626833	2.415636	0.3215	0.1073
2	2.464400	2.378610	0.3274	0.1096
3	2.385800	2.369173	0.3283	0.1101

ROUGE-L	ROUGE-L SUM	BLEU	Gen Len
0.1955	0.1955	0.3361	114.6246
0.1996	0.1995	0.3560	101.8699
0.2002	0.2001	0.3584	100.1943

In general, the three fine-tuned models, BART, PEGASUS and ProphetNet, were able to generate relatively good summaries. These were mostly accurate and coherent, capturing the key points from the original texts. However, there were some limitations and anomalies in their performances. BART performed well in generating summaries, however, in most cases, the summaries were incomplete. Many summaries were accurate, relevant and coherent, however, some had clarity and conciseness issues, with some missing the final sentences or repeating the same phrase multiple times. Pegasus performed well, producing summaries that were accurate, relevant and coherent, however, similarly to BART, some summaries were incomplete, missing the final sentences, and sentences repeated multiple times; this directly affected the clarity and conciseness of the summaries. ProphetNet performed similarly to BART and PEGASUS, the summaries were accurate, but the completeness and conciseness issues persisted, the final sentences were missing in some summaries, and sentences were repeated multiple times. Despite this, the summaries were clear and easy to understand. Some of the anomalies found were:

- Truncated texts: all three models suffered these issues; many generated summaries were missing the last sentence, which affected their completeness and clarity.
- Repetitions: some summaries repeated the same concepts multiple times; this was most noticeable with PEGASUS.
- Ignored details: in some cases, the three model summaries were missing key details, which affected their completeness.
- Misinformation: in one case, with ProphetNet, the model introduced false information which was not present in the original text.

In contrast, and as expected, the summaries generated by the untrained GPT-4 were fully accurate, complete, relevant, coherent, concise and clear.

6 Conclusions

The research aim was to fine-tune the three Large Language Models, BART, PEGASUS and ProphetNet, on three different datasets each, and subsequently evaluate their performance in generating new summaries. The results showed different levels of performance across the three models, with their strengths and limitations, and initially, the results were disheartening, the ROUGE and BLEU scores for the untrained BART model were much higher than expected; however, the results normalized with PEGASUS and ProphetNet, but still the difference between the untrained and the finetuned models was negligible to have any substantial effect on the final results. The generated summaries were accurate, relevant and coherent; however, they had completeness and conciseness issues due to missing final and repeated sentences. Notable, in one instance ProphetNet introduced false information, information that was not present in the original text. Comparatively, the zero-shot GPT-4 generated summaries were almost perfect in each section. This is not surprising due to the sheer number of parameters, 1.7 trillion, almost 1000 times more parameters than all three fine-tuned models combined. This indicated that general Large Language Models, on the scale of billion and trillion parameters, can perform certain tasks without needing to be specially fine-tuned. However, fine-tuning should not be ignored, An OpenAI article [17] explains the potential benefits of fine-tuning GPT models on custom data; some of the use cases they provide are improved steerability, reliable output formatting, and regulating custom tone. This research demonstrated the efficacy of fine-tuning LLMs, its impact is limited by the size of the training data and model; in comparison, larger models, like GPT or PaLM, can easily outperform them; this indicates that the number of parameters is key. Finally, the research has set the stage for future investigations into optimising fine-tuning for small to mid-sized Large Language Models to achieve performance similar to larger models. Another question is how to efficiently deploy these models given the limited amount of resources, as these models require high computational power and therefore have a high carbon footprint, and how to sustainably develop and deploy newer models. Energy considerations, as investigated in [22], affirm that training a model with 213 million parameters releases 284,019 KG carbon dioxide CO₂ in the atmosphere, the equivalent of what 5 cars release in their lifetime.

References

1. Adams, P. Kerch bridge is hated symbol of Russian occupation. BBC News, 07 2023.
2. Chu, J. The ocean's color is changing as a consequence of climate change: The color changes reflect significant shifts in essential marine ecosystems., 07 2023.
3. Fabbri, A. R., Li, I., She, T., Li, S., and Radev, D. R. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model, 06 2019.
4. Ghadimi, A. and Beigy, H. Sgcsomm: An extractive multi-document summarization method based on pre-trained language model, submodularity, and graph convolutional neural networks. *Expert Systems with Applications*, 215:119308, 04 2023.
5. Goldsack, T., Zhang, Z., Lin, C., and Scarton, C. Making science simple: Corpora for the lay summarisation of scientific literature, 10 2022.
6. Gov.uk. Civil aviation authority review: report, 07 2023.
7. Gov.uk. Higher education policy statement and reform government consultation response, 07 2023.

8. Gov.uk. HMRC's annual report and accounts 2022 to 2023: performance overview, 07 2023.
9. Huang, L., Cao, S., Parulian, N. N., Ji, H. C., and Wang, L. Efficient attentions for long document summarization. *arXiv*, 04 2021.
10. HuggingFace. Summarization.
11. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv (Cornell University)*, 10 2019.
12. Li, Y., Sha, L., Yan, L., Lin, J., Rakovic, M., Galbraith, K., Lyons, K., Gašević, D., and Chen, G. Can large language models write reflectively. *Computers and Education: Artificial Intelligence*, 4:100140, 01 2023.
13. Mackenzie, J. South korea floods: Tunnel horror brings home climate fears. *BBC News*, 07 2023.
14. Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., and Sundheim, B. Summac: a text summarization evaluation. *Natural Language Engineering*, 8:43–68, 03 2002.
15. Mridha, M. F., Lima, A. A., Nur, K., Das, S. C., Hasan, M., and Kabir, M. M. A survey of automatic text summarization: Progress, process and challenges. *IEEE Access*, 9:156043–156070, 2021.
16. Orasan, C. Automatic summarisation: 25 years on. *Natural Language Engineering*, 25:735–751, 09 2019.
17. Peng, A., Wu, M., Allard, J., Kilpatrick, L., and Heidel, S. Gpt-3.5 turbo fine-tuning and api updates, 08 2023.
18. Pohl, J. Deep-sleep brain waves predict blood sugar control, 07 2023.
19. Qi, W., Yu, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., and Zhou, M. Prophetnet: Predicting future n-gram for sequence-to-sequence pretraining. *arXiv*, 01 2020.
20. Stallard, E. and Poynting, M. Extreme weather: What is it and how is it connected to climate change? *BBC News*, 08 2021.
21. Stella, F., Della Santina, C., and Hughes, J. ChatGPT designs a robot, 06 2023.
22. Strubell, E., Ganesh, A., and McCallum, A. Energy and policy considerations for deep learning in nlp. *arXiv:1906.02243 [cs]*, 06 2019.
23. Wu, N., Gong, M., Shou, L., Liang, S., and Jiang, D. Large language models are diverse role-players for summarization evaluation. *arXiv*, 03 2023.
24. Yang, A. K., Liu, K., Liu, J., Lyu, Y., and Li, S. Adaptations of rouge and bleu to better evaluate machine reading comprehension task. *Proceedings of the Workshop on Machine Reading for Question Answering*, 01 2018.
25. Zhang, J., Zhao, Y., Saleh, M., and Liu, P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv*, 12 2019.
26. Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. Benchmarking large language models for news summarization, 01, 2023.