# Exploring the Summarization Landscape: A Comparative Study of GPT-2, T5, and GPT-3.5 in NLP

**Anonymous ACL submission**

## Abstract

This study investigates the field of automatic text summarization within Natural Language Processing (NLP), focusing on the performance of three advanced deep learning models: GPT-2, T5, and GPT-3.5, using the CNN/Daily Mail dataset. Each model, with its distinct architectural design, is rigorously assessed using a variety of metrics, including ROUGE scores, BERT scores, and manual evaluation methods. The study not only compares the models' performances but also provides insights into the strengths and limitations of different architectural approaches. Our findings reveal that each model excels in different facets of summarization, highlighting how diverse architectures impact performance. Emphasizing the need for varied evaluation techniques, this study enhances our understanding of NLP models' capabilities and paves the way for future advancements in the field of automatic text summarization.

## 1 Introduction

In the modern era, characterized by a rapid influx of information, automatic text summarization has emerged as a key area in Natural Language Processing (NLP). This technology, designed to condense original text into accurate and coherent summaries, plays a crucial role in facilitating quick understanding of large volumes of data. Its importance is particularly pronounced in domains like news reporting and online content, where timely and succinct information is vital.

Our study is propelled by the objective to explore and benchmark the capabilities of advanced deep learning models in the field of text summarization. We focus on three prominent models: the Generative Pretrained Transformer 2 (GPT-2) small model, the Text-to-Text Transfer Transformer (T5) base model, and the Generative Pretrained Transformer 3.5 (GPT-3.5). Each of these models brings a unique approach to handling NLP tasks, offering rich potential for comparative analysis. GPT-2 and GPT-3.5, both products of OpenAI, have been recognized for their broad application and effectiveness in various NLP challenges. In contrast, Google's T5 employs a distinct text-to-text format, offering a valuable comparative perspective.

We hypothesize that among these models, GPT-3.5, with its advanced architectural design, will stand out in terms of effectiveness in automatic text summarization. To test this hypothesis, our research utilizes the CNN/Daily Mail dataset, accompanied by a comprehensive array of evaluation metrics including ROUGE scores, BERT scores, and human assessments. These tools collectively form an integrated framework, allowing for a thorough assessment of both the quantitative and qualitative dimensions of each model's summarization capabilities.

Aiming to offer in-depth insights into the field of automatic text summarization, this study not only evaluates current model performances but also seeks to contribute to future research directions. It underscores the significance of varying model architectures in the realm of natural language processing, setting a foundation for future explorations into how these differences impact task performance in NLP.

## 2 Related works

Automatic text summarization, a key area in Natural Language Processing (NLP), has undergone significant evolution since its early days. The initial phase focused on extraction-based methods, which were predominantly rule-based and aimed at identifying key phrases and sentences from texts without altering the original content (Luhn, 1958; Edmundson, 1969). These methods laid the groundwork for the field but were limited in their ability to generate contextually rich summaries.

The advent of deep learning technologies

marked a transformative shift in text summarization strategies. The field gradually moved towards abstractive summarization methods, a more advanced approach that involves creating new text segments to encapsulate the essence of the original content. This shift necessitated a deeper level of natural language understanding and processing capabilities, as demonstrated in seminal works like that of Rush et al., 2015 and See et al., 2017.

Recent advancements have been particularly influenced by the development and application of transformer-based models, such as GPT-2, T5, and GPT-3.5. These models have set new standards in the quality of generated summaries. GPT-2, a creation of OpenAI, illustrated the potential of transformer architectures in producing coherent and contextually relevant text, making it a suitable candidate for summarization tasks (Radford et al., 2019). Google's T5 model, with its innovative text-to-text approach, treats all NLP problems as text-generation tasks and has shown promising results in summarization (Raffel et al., 2020). The introduction of GPT-3.5 represented a significant advancement, offering more nuanced and accurate text generation due to its increased scale and capability (Brown et al., 2020).

The evaluation of these models often employs datasets like CNN/Daily Mail to benchmark their performance in summarization. Standard metrics used in these evaluations include ROUGE scores, which measure the overlap between the generated and reference summaries, and BERT-based evaluations for semantic similarity assessment. The latter, introduced by Devlin et al., 2018, has added a new dimension to how semantic understanding is quantified in NLP models (Lin, 2004; Zhang et al., 2019).

# 3 Method

## 3.1 Dataset selection

The CNN/Daily Mail dataset is a comprehensive collection of news articles from CNN and Daily Mail accompanied by summaries. This dataset is renowned for its size and diversity, making it a prime choice for text summarization research. Given the constraints imposed by our model architectures, particularly GPT-2 and T5, we opted for articles with a word count below 768. A balance between computational efficiency and content richness guided this decision.

We meticulously curated a representative subset from the complete dataset. This subset was respited into training, validation, and test sets. It is important to note that GPT-2 and T5 models were trained and validated on the same training (validation) set and evaluated on a standard test set, fostering a consistent, comparable framework. Similarly, GPT-3.5, although not trained on this dataset, was also evaluated using the same test set to maintain uniformity in assessment.

## 3.2 Model selection and experiments

Our computational resources dictated a selective approach to model training. We chose to fine-tune GPT-2 and T5 on smaller batches of data. The fine-tuning process was carefully monitored to optimize model performance while preventing overfitting.

We fine-tuned the GPT-2 small model on one Nvidia RTX3060 6GB GPU. Given the constraints of the GPU's capacity, we limited ourselves to a batch size of one. We also accumulated gradients across 32 steps before updating the model's weights. This approach allowed us to train the large model efficiently with the available resources. While training, we concatenated sources (summaries) and targets (articles) in training examples with a separator token (<|sep|>), a delimiter in between, padded with the padding token (<|pad|>), and another delimiter, up to a context size of 1024 for GPT-2, respectively. Additionally, we used a cross-entropy loss function and excluded the loss from padding tokens to enhance the quality of the summaries. For the learning rate, we combined warmup and one-cycle policies to optimize the training process. We experimented with various temperature and beam width settings during the summary generation phase for nucleus sampling and beam search. Details of the specific hyperparameter settings are provided later.

We fine-tuned the T5 base model on one Nvidia RTX3070 8GB GPU. We refer to Google's official description of the T5 model to adapt the training data format, including two columns, "source_text" and "target_text", which correspond to the original news article and the Gold Standard summary, respectively. We also added the task prefix "summarize:" for each news article according to the requirements. When fine-tuning, the T5 model also applies to the maximum news length of 768 tokens, the excess will be truncated. Similarly to the GPT-2 model, we employed padding tokens (<|pad|>) when processing the inputs of the T5 model to

| Hyperparam. | GPT-2 | T5 |
|---|---|---|
| Learning Rate | 1e-5 | 1e-4 |
| Batch Size | 1 | 2 |
| Grad accumulation | 32 | / |
| Max Generate Length | 100 | 100 |
| Beam Number | 3 | 4 |
| Top K | 30 | 50 |
| Top P | 0.75 | 0.95 |
| Temperature | 0.8 | 1 |
| Repetition Penalty | / | 2.5 |
| Length Penalty | / | 0.6 |

Table 1: List of optimal hyperparameters for two controllable models after fine tuning

padding the inputs with less than 768 tokens to 768 tokens. Notably, to simplify the code implementation, we chose the PyTorch-lighting library, rather than the vanilla PyTorch library, to implement the fine-tuning architecture for the T5 model. When tuning the hyperparameters, some of the hyperparameters are fixed according to the original paper of the T5 model, such as the number of beams during beam search, and extensive experiments are conducted on the rest of the hyperparameters. For better results, we adopted CosineAnnealingLR and carefully tuned its parameters, and we also introduced an early stopping mechanism after the third epoch to prevent overfitting. Table 1 shows the training and prediction hyperparameters for both our GPT-2 and T5 models.

Finally, for comparison purposes, we evaluated GPT-3.5 using a zero-shot approach through ChatGPT. We selected the first 30 articles from our test set and applied prompt engineering to guide the summarization process without prior training on our dataset. Noticing GPT-3.5's inclination towards longer outputs, which could potentially bias our evaluation methods, such as coverage of highlights, we provided specific instructions in our prompts to ChatGPT to generate summaries with an approximate length of 47 words. This length corresponds to the average of our ground truth summaries, ensuring a more equitable comparison.

### 3.3 Evaluation

The evaluation of the models' summarization capabilities was conducted using several metrics:

1. ROUGE Score: The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score measures the overlap of n-grams between the generated summary and a set of reference summaries. This metric is pivotal in gauging how much a model captures critical information from the original text.

2. BERT Score: To assess the semantic quality of the summaries, we employed the BERT score, which evaluates the cosine similarity of token embeddings between the generated and reference texts. This score is instrumental in understanding the depth of contextual comprehension by the models.

3. Manual Evaluation: To supplement our automated metrics, we conducted manual evaluations focusing on information coverage and fluency of the generated summaries. This crucial step involved human evaluators who meticulously analyzed each summary, assessing how well the models captured the essential elements of the original texts (coverage of highlights), and the overall readability and coherence of the summaries. The 'Coverage of Highlights' method evaluates the effectiveness of the model-generated summaries in encapsulating key points or highlights from the original text. Additionally, the 'Fluency Rating' assesses the readability and linguistic quality of the summaries on a scale from 1 to 5, focusing on grammar, syntax, and overall coherence, with ratings ranging from poor (1) to excellent (5).

## 4 Results

Tables 2 and Table 3 show the numerical results of the models on different test corpora. Note that given the constraints of time and assessment resources, only 30 articles were used for the human evaluation. The value in parentheses is the difference between the score from the model and the score from the baseline method (usually GPT-2) under the same criteria.

See Figure 1 below for examples of summaries generated by the three models. Each frame represents a news item in which the body of the news item is omitted and only its original summary, i.e., the ground truth summary is retained. Both summary examples coincidentally comprise three key nuggets of information, marked by 1, 2, 3, which are also anticipated to be captured in the generated summaries. Beneath the Ground Truth Summary

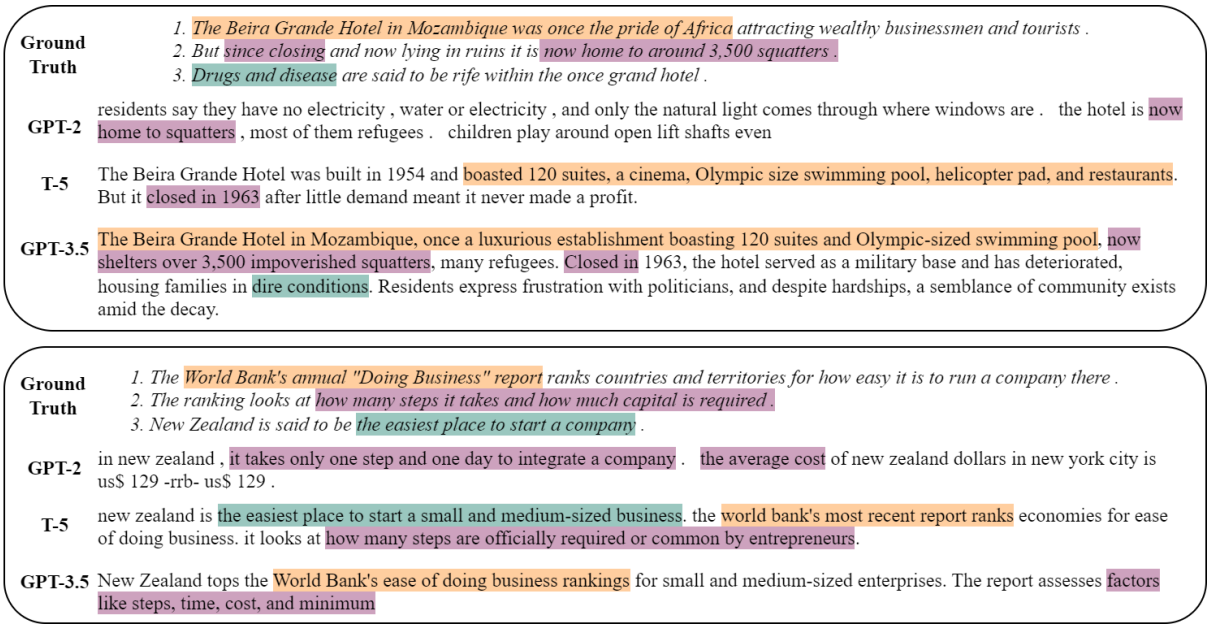


Figure 1: Two sample news clips, including a comparison of the ground truth summary with the summary generated by the three models. Where different colored text backgrounds represent key information points.

are summaries generated by each of the three models, where text in a background color indicates that it captures a key information point in the corresponding Ground Truth Summary. In particular, the length of the output summaries of both GPT-2 and T5 models is limited to a maximum of 100 tokens, whereas GPT-3.5 does not have such a hard programmatic limit.

In a manual evaluation of the text summarization capabilities of GPT-2, T5, and GPT-3.5 models, we observed significant disparities in capturing highlight information. Taking the Beira Grande Hotel summary as a case study, GPT-2 could only capture a singular key point, highlighting the hotel's current status as a squatters' residence. This suggests GPT-2's capacity to identify present-state information, yet potentially overlooking the complexity of historical context and situational depth.

In contrast, the T5 model identified two key points: the hotel's past opulence and its year of closure, indicating that the model can retain more key points when generating summaries. However, GPT-3.5 demonstrated superior information coverage, integrating not only the aspects identified by T5 but also elaborating on the current dire circumstances of the inhabitants, indicating an in-depth understanding of the textual nuances. Concerning fluency, the narrative quality and coherence of the models' outputs varied, with GPT-3.5's summaries exhibiting notable linguistic naturalness and logical flow, likely due to its advanced algorithms and extensive training data corpus. While GPT-2 and T5 also produced relatively coherent text, they occasionally fell short of the narrative fluidity displayed by GPT-3.5.

This dataset exemplifies the varying efficacy of models in automated text summarization, providing a clear benchmark for future research in selecting the most appropriate model for specific summarization tasks.

## 5 Discussion and conclusion

The study's results reveal a distinct performance ranking among the models. While GPT-2 demonstrated specific capabilities, it was outperformed in text summarization tasks by both T5 and GPT-3.5. In automated metrics, T5 slightly led over GPT-3.5 with ROUGE scores (ROUGE-1: 0.354 vs. 0.346 for GPT-3.5) and BERTScore (0.865 vs. 0.854), indicating a marginal but notable advantage. Conversely, given its massive amount of pre-training data, GPT-3.5 excelled in the manual evaluations, particularly in coverage of highlights and fluency ratings (58.82% coverage and a fluency rating of 4.0), underscoring its superior performance in these areas.

While ROUGE and BERTScore provide valuable quantitative assessments, they typically do not fully capture the deep semantic understanding and nuance of text as perceived by humans.

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | BERTScore |
|---|---|---|---|---|
| GPT-2 | 0.279 | 0.078 | 0.260 | 0.821 |
| T5 | 0.354 (+0.075) | 0.156 (+0.078) | 0.332 (+0.072) | 0.865 |

Table 2: Comparison of the fine tuned models on the entire test set (GPT-2, T5)

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | BERTScore | Coverage (%) | Fluency (1-5) |
|---|---|---|---|---|---|---|
| GPT-2 | 0.267 | 0.059 | 0.241 | 0.819 | 33.33 | 2.83 |
| T5 | 0.355 (+0.088) | 0.152 (+0.093) | 0.258 (+0.017) | 0.859 (+0.040) | 51.11 (+17.78) | 3.5 (+0.67) |
| GPT-3.5 | 0.346 (+0.079) | 0.114 (+0.055) | 0.226 (-0.015) | 0.854 (+0.035) | 58.82 (+25.49) | 4.0 (+1.17) |

Table 3: Comparison of the all the 3 models on the first 30 test set articles (GPT-2, T5 and GPT-3.5)

Human evaluations, considering factors like coherence, relevance, and overall readability, are critical in assessing summarization models' actual practicality and effectiveness in real-world applications. While the overall results are in line with our expectations, the marginally better performance of T5 in ROUGE and BERTScore metrics presents a notable exception, making it an intriguing point of discussion. The T5 model's architecture, designed for efficiency in tasks like summarization, coupled with its fine-tuning on our dataset, likely contributed to its alignment with our task's specific content and style requirements. In contrast, GPT-3.5, though not trained on our dataset, demonstrated its versatility and advanced capabilities. Comparing to T5, GPT-3.5 got slightly lower scores in ROUGE and BERTScore, underscoring the significance of model-specific training on task-specific datasets for achieving high scores in specific automated metrics.

## 6 Statement of contributions

This study was conceptualized by Haoyue, Linrui, and Yujun. Haoyue and Linrui selected and prepared the dataset. Haoyue led the GPT-2 modeling and experiments, while Linrui handled the T5 model. Yujun was responsible for evaluating the GPT-3.5 model, including analysis and comparison of all model outputs, integrating manual evaluation results. The initial draft of the manuscript was written by Haoyue. All three contributors, Haoyue, Linrui, and Yujun, actively participated in revising their respective sections of the manuscript and collaboratively worked towards the completion of the final report.

## 7 Code repository

We published the code to github for reference and reproduction: github.com/haooyuee/Generating_Text_Summary_GPT2

## References


Tom B Brown et al. 2020. Language models are few-shot learners.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Harold P Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

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

H P Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Alec Radford et al. 2019. Language models are unsupervised multitask learners.

Colin Raffel et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Tianyi Zhang, Junjie Zhao, and Yann LeCun. 2019. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR*.