CSE 561 Project Final Report: Legal Document Summarization

Authors: Sterling Lech, Jake Valentine, Yiding Chen
Washington University in St. Louis
1 Brookings Dr, St. Louis, MO 63130
l.sterling@wustl.edu, j.a.valentine@wustl.edu, c.yiding@wustl.edu

1. ABSTRACT

Our project successfully creates a framework for evaluating text summarization using large language models (LLMs) and uses it in combination with other metrics to evaluate the performance of our own models trained to summarize legal documents. Our framework evaluates non-reference-based summarization metrics: coherence, fluency, and consistency, reducing reliance on biased and time-intensive human judgments. These metrics are combined with reference-based scores like ROUGE-L and BERT to assess precision and recall. Our results indicate that our evaluation framework can be used to replace human evaluators demonstrate the effectiveness of LLMs in automated summary evaluation. The evaluation framework allows us to find the shortfalls of our legal document summarization models and areas where they are improving upon the baseline model. Our first trained model shows decreased performance in precision and recall, but as highlighted by our evaluation framework, shows improvements in summary consistency. Our second trained model shows improvements when using traditional evaluations with ROUGE-L score and BERT score, but appears to overfit the training data when analyzed with our evaluation framework. Overall, this project helps to narrow the gap between automated and human summary evaluation and makes important strides in optimizing legal document summarization.

2. INTRODUCTION

Reading full legal documents can be a challenging task; however, their contents are important to legal professionals as well as their clients. Legal documents are often lengthy and filled with formal language and jargon that may be difficult for readers to understand. Nonetheless, having a comprehensive understanding of a legal document is crucial for individuals in legal settings as well as non-legal settings such as businesses. Gaining a quick and thorough understanding of legal documents has many benefits in addition to saving time. It reduces reliance on costly legal professionals and can reduce the potential for legal conflict as all parties are more likely to have an understanding of any legal matter pertaining to them. For example, a business will know exactly what terms they agreed to in a license agreement that they may have otherwise never read or understood. In this project we train a large language model for legal document summarization and evaluate it using a novel framework developed for automatic text summarization using a variety of performance metrics.We aim to improve upon existing models in terms of legal summarization performance and build upon previous evaluation metrics with our evaluation framework.

3. RELATED WORK

Text-to-summarization evaluation: Following the development of reference-based metrics like ROUGE [1], which measures n-gram overlap, more advanced semantic evaluation methods, such as BERTScore [2], were introduced. BERTScore leverages contextual embeddings to assess similarity, addressing the limitations of ROUGE in handling paraphrased or lexically diverse text.

Legal text summarization: Legal summarization has garnered attention due to its unique challenges, including domain-specific language and structure. Deroy et al. [3] investigated the use of large language models for summarizing legal judgments, demonstrating their potential in this domain. Bhattacharya et al. [4] performed a comparative study of summarization algorithms for legal case judgments, highlighting the need for customized approaches.

Domain-specific model advancements: Nigam and Deroy [5] explored fact-based court judgment prediction, emphasizing the need for factual correctness in legal summarization models. Their work underscores the importance of integrating domain-specific constraints into text generation tasks.

4. EVALUATION FRAMEWORK

We propose a novel framework for evaluating generated summaries using pre-existing large language models. Since most prior work in summary evaluation deals with reference-based metrics that compare a ground truth reference summary to the generated summary such as precision and recall, evaluating other important aspects of summaries such as fluency is challenging. While human judgment can be used to evaluate these metrics, this is impractical as it would require someone to spend a long period reading summaries and is also less reliable due to human biases and inconsistencies. Consequently, our framework aims to use large language models to bridge the gap between human and automatic evaluation of non-reference-based metrics–coherence, fluency, and consistency. We conduct experiments to understand the feasibility of using a large language model to replace a human evaluator and to find the optimal prompting strategy for each evaluation metric.

4.1 EXPERIMENTS

We conduct our experiments using the SummEval dataset [7], which contains 16000 samples of generated summaries with labels scoring each summary on coherence, fluency, and consistency. Each metric is rated on an integer scale from 1 to 5 by 8 humans (5 crowdsourcers and 3 experts) and the average is provided in the dataset. For our experiments, we randomly select 100 samples. While it would have been useful to utilize a larger portion of the available data, we restricted our sample size due to a limited budget for making external API requests to pre-existing large language models. However, we believe our sample size of 100 summaries is sufficient for developing the framework and will produce results that can be generalized. Additionally, this dataset is not specifically catered towards legal document summarization as the provided summaries are from CNN and Daily Mail articles.

The goal of our experiments is to develop a prompting strategy that produces responses that best align with human judgments. As such, we test multiple prompting strategies on two different

large language models: GPT-3.5 Turbo and GPT-4o. These models were selected to represent different levels of model complexity while being relatively cheap to use and easy to access. On each model, we test two prompt types: guided prompts and unguided prompts. The guided prompts indicate the criteria for a summary to be ranked at each number 1 through 5 whereas unguided prompts allow the model to have free judgment and develop its own scoring system as would a human annotator. Both prompts define the same base task, so the relationship between the unguided and guided prompt is defined as *guided prompt = unguided prompt + specific scoring criteria*. *Figure 1* shows the guided prompt used to evaluate coherence. It contains the unguided prompt which instructs the model to rate the coherence within the full range of 1 through 5 and additionally guides the model on how to score the summary using brief descriptions. Each prompt was used for a single evaluation and for an aggregated evaluation found by averaging the scores over 3 requests resulting in 4 different prompting strategies. Each strategy is tested on both GPT models and the correlation between human evaluations and LLM-generated evaluations is found.

```
Please rate the coherence of the following summary on a scale from 1 to 5, where:
1 - Very poor coherence, difficult to follow but still somewhat understandable.
2 - Poor coherence, hard to follow, but some parts are still clear.
3 - Fairly coherent, but contains some unclear sections.
4 - Mostly coherent with small issues or ambiguities.
5 - Very coherent, clear and easy to understand.
Do not hesitate to score at the extremes.
Provide only the score and nothing else.
```

(Figure 1: guided prompt used to evaluate coherence)

4.2 RESULTS AND DISCUSSION

By definition, coherence measures how logically connected and sensible the response is. Fluency measures how natural, grammatical, and human-like the responses are. Finally, consistency measures whether the model provides reliable and non-contradictory information. All of these metrics are crucial for legal document summarization as the information must be accurate, easy to understand, and professional. Figure 2 shows the correlation between the human judgments and LLM-generated evaluations for each model and prompting strategy.

		Coherence		
	Unguided	Unguided 3x	Guided	Guided 3x
GPT3.5	0.436	0.500	0.209	0.386
GPT4	0.519	0.609	0.545	0.577
		Fluency		
	Unguided	Unguided 3x	Guided	Guided 3x
GPT3.5	0.312	0.289	0.297	0.268
GPT4	0.524	0.590	0.421	0.510
		Consistency		
	Unguided	Unguided 3x	Guided	Guided 3x
GPT3.5	0.255	0.259	0.275	0.278
GPT4	0.440	0.471	0.511	0.549

(Figure 2: correlation between the human judgments and LLM-generated evaluations for each model and prompting strategy.)

LLM-generated evaluations correlate positively with human evaluations for all tested prompts and models; however, for most prompting strategies and metrics, the evaluations generated by GPT-3.5 Turbo correlate poorly (between .2 and .4). Those generated by GPT-4 provide stronger correlations, typically between .4 and .6. This indicates that GPT-4 can evaluate summaries with decent similarity to human judgments. While LLMs and humans may score certain summaries differently, correlation is a useful indicator of success because it measures the rankings of summaries relative to each other, meaning summaries that score better in human evaluations also score better on LLM evaluations. The results also show that for almost all prompting strategies, there is a significant improvement in correlation when going from a single evaluation to an average of over 3 evaluations. This indicates that, like humans, LLM evaluations are not deterministic and have variation when evaluating a summary multiple times. Also, averaging multiple evaluations allows for more potential scores which allows the distribution of the scores to better align with the human evaluations. For example, using 3 evaluations allows the framework to differentiate between a summary that scores (5, 5, 5) from one that scores (5, 4, 5). For coherence and fluency, we see the best results when using unguided prompts, whereas consistency correlates the best when using guided prompts. This suggests that GPT-4 has a better understanding of what makes a summary coherent or fluent compared to consistent and benefits from being given clear guidelines for ranking consistency. The simple instructions provided in the guided prompt may cause the model to be less creative in its evaluations causing it to have worse evaluations for coherence and fluency. Overall, our results are promising and show the plausibility of using a large language model to replace human evaluators to score generated summaries. While GPT-3.5 Turbo doesn't show significant correlations, those found when using GPT-4o are semi-strong and are likely to improve as large language models continue to improve.

5. LEGAL DOCUMENT SUMMARIZATION MODELS

5.1 TRAINING

To train a model to summarize the legal documents, we first import an untrained general text summarizing model from the Hugging face: Legal_Text_Summarization-lama2 created by AjayMukundS. Then we utilized a curated dataset of approximately 8,000 rows containing legal case judgments and their corresponding concise summaries. This dataset is in the Parquet format, preprocessed by tokenizing the input text (full judgments) with padding, truncation, and a maximum length of 1024 tokens, while the target summaries were tokenized similarly but capped at 256 tokens. Two training runs were conducted with variations in configuration to optimize performance. Because those runs will occupy an enormous amount of computational resources, we rented RunPod GPU Cloud service to finish our training. In Run 1, we used a batch size of 16 with BF16 mixed precision, gradient checkpointing for memory efficiency, and Hugging Face's AdamW optimizer with a weight decay of 0.02. The training spanned 10 epochs with a learning rate of 3e-5, a cosine scheduler, and checkpoints saved every 1,000 steps. In Run 2, we refined the setup with an effective batch size of 32 (via gradient accumulation), FP16 precision, and a higher learning rate of 5e-5 using a linear scheduler. More frequent evaluations were conducted every 250 steps with text generation enabled, and checkpoints were saved every 500 steps. While both runs achieved improved training losses (1.5169 and 1.3840, respectively), the evaluation loss remained comparable (around 1.9), indicating the potential for further tuning to enhance generalization.

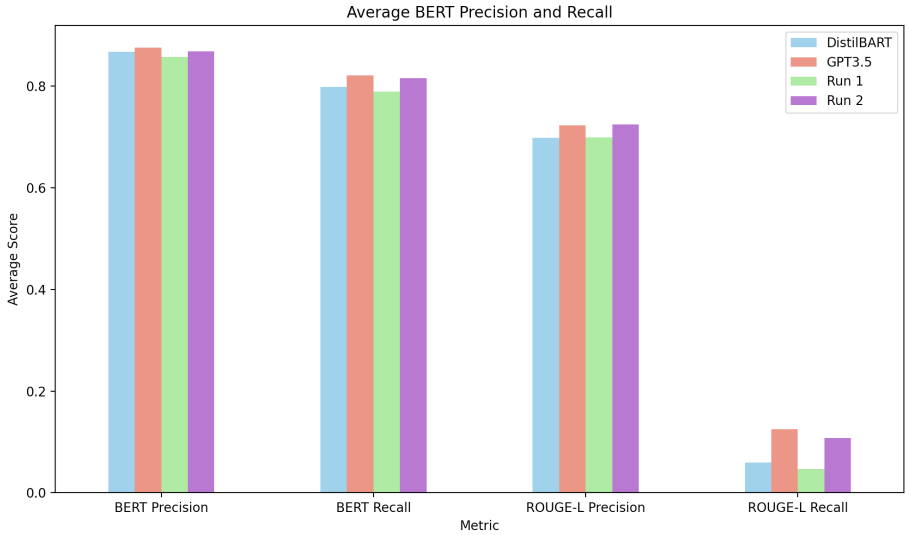
5.2 EVALUATION

Since there is no single way to evaluate a summary, we evaluate our models using a collection of metrics. We combine non-reference-based metrics coherence, fluency, and consistency with reference-based metrics precision and recall to assess the performance of our legal document summarization models. Precision measures how much of the generated summary appears in the reference summary, which evaluates how important the text in the generated summary is. Recall measures the amount of the reference summary that is found in the generated summary. This tells us how complete the generated summary is in containing all of the important information of the reference summary. For each of our models, we generate 200 summaries of legal documents from our testing dataset. We find precision and recall using two different automatic scoring methods: ROUGE-L score and BERT score. Furthermore, following the framework outlined in section 4, we find the coherence, fluency, and consistency of each summary.

ROUGE-L score (Recall-Oriented Understudy for Gisting Evaluation) [1]. ROUGE-L score measures the amount of overlap between the reference summary and the summary generated by the model by finding the longest common subsequence (LCS) between the two summaries. We evaluate our model using two ROUGE-L metrics: precision and recall. Additionally, we evaluated the model using the BERT score [2]. BERT score is another reference-based evaluation method that calculates the pairwise cosine similarity between each token in the reference summary and generated summary. Contextual embeddings for the summaries are created using the BERT language model. BERT score addresses some of the shortfalls of the ROUGE score. While the ROUGE score looks for exact matches between tokens, the BERT score measures similarity making it suitable to handle paraphrasing which is beneficial for evaluating the output of our model which is likely to contain lexical variation. As with ROUGE-L, we find precision and recall.

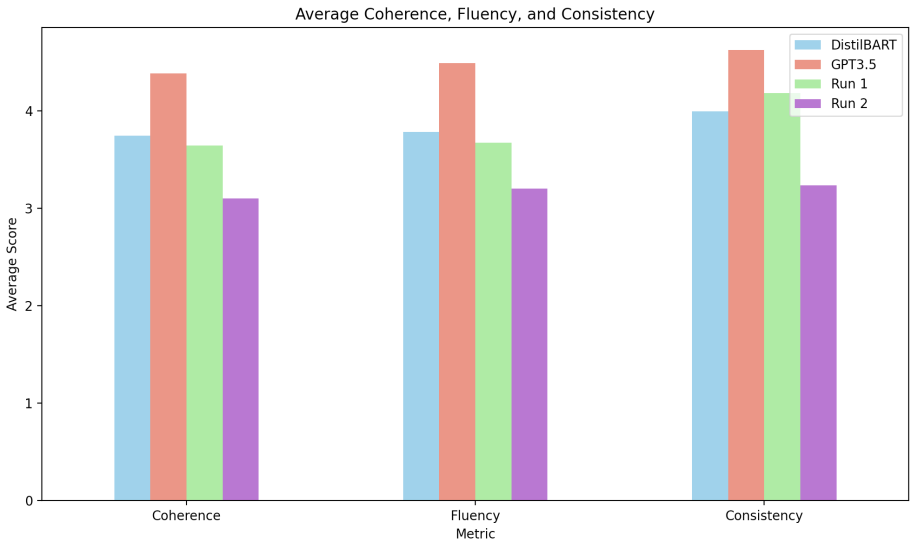
5.3 RESULTS AND DISCUSSION

Figure 3 shows the average precision and recall calculated using BERT score and ROUGE-L score for our trained models (run 1 and run 2), the baseline DistilBART model, and GPT-3.5.



(Figure 3: Average precision and recall comparison)

Our first model slightly underperforms DistilBART in all metrics. While the model still performs well in most metrics, it does not show an improvement in precision and recall compared to the baseline model, which indicates that it did not fit very well to the training data. On the other hand, our second model outperforms DistilBART in all metrics and performs almost as well as GPT-3.5. This improvement indicates that our second model includes more relevant legal information and includes less information that does not match the reference summary. This is especially consistent with the major increase in ROUGE-L recall which explains that the model is incorporating more direct legal language that is found in the documents. This indicates that our second training process allowed the model to better fit to the training data which resulted in improved performance.



(Figure 4: The average score of coherence, fluency, and consistency in different training attempts)

Our analysis of the results continues in *Figure 3* which shows the bar chart comparing the average scores of coherence, fluency, and consistency for the same models. However, unlike with precision and recall, we see a considerable decrease in coherence, fluency, and consistency for our second model. While the model shows improvements in its ability to summarize legal documents to contain relevant information it loses some of its general abilities to write quality English text. This result is not extremely surprising because legal documents contain text that is often confusing and doesn't read like regular English, and as previously established, this model incorporates more direct legal text into its summaries. GPT-3.5 though, can include more relevant legal text while still leading in coherence, fluency, and consistency over the other models. This all may suggest that our second model is overfitting the training data and not generalizing well to legal documents in general. However, it is unrealistic to expect our models to perform as well as GPT-3.5 since they are not comparable in size. Our first model shows a slight decrease in coherence and fluency, but an improvement in consistency over DistilBART. This indicates that the model is better at understanding complicated legal texts and presents summaries with fewer inaccuracies and contradictions than the other models.

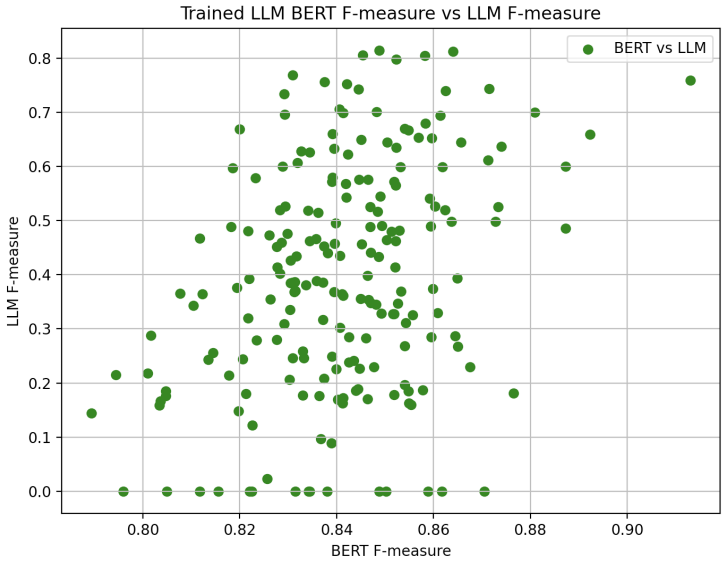
While the numerical results of our model training show some improvements in the ability to summarize legal documents, our results highlight the usefulness and impact of our LLM-based evaluation framework. Using traditional automatic evaluation methods like ROUGE score and BERT are supplemented by our evaluation framework to gain more complete insights into the performance of the models. For example, without the evaluations on coherence, fluency, and consistency, it appears that our first model is purely worse than DistilBART and our second model is purely better. Being able to easily generate other reliable and important evaluation metrics allows us to understand that our first model has improvements and our second model has downsides when compared to the baseline model. Our framework helps to bridge the gap between the automated summary evaluation and human evaluation which is an important step in optimizing language models for legal document summarization and summarization in general.

6. LIMITATIONS AND FUTURE WORK

The primary limitation we encountered when developing our summary evaluation framework was the price of prompting external large language model APIs. While GPT-3.5 Turbo was very cheap, our results indicated that GPT-4o was overwhelmingly more useful as an evaluator which was a much more costly model to prompt. As a result, our experiments to determine the optimal prompts had to be simplified to fit our cost constraints. Our dataset contained 16000 summaries and their corresponding evaluations, of which we only used 100. Had we had a larger budget for this project, we could have utilized a greater portion of this data which would have allowed our results to be more convincing. Furthermore, it would have been beneficial to try more prompt variations and strategies. In particular, one useful strategy would have been to use one-shot or even multi-shot prompting by providing the model example summaries and human evaluations before asking for an evaluation of a new prompt. We would have expected to see stronger results using this strategy, however; the price would have increased drastically if we had to input entire summaries into the prompt for each evaluation. This would be a great next step for improving the evaluation framework in the future. We also would have liked to develop our evaluation

framework so it could be the sole framework for generated summaries rather than relying on a combination of other metrics such as the ROUGE-L score and BERT score. We did preliminary investigations into using an LLM to find reference-based metrics; however, due to the cost of inputting reference summaries into the prompt, we restricted our exploration to a single prompt on GPT-3.5 Turbo which was not a sufficient experiment to include in the findings of this paper. As shown in *Figure 5* which shows a plot of BERT score and an LLM-given score to measure precision and recall, there appears to be a noticeable relationship that could be further explored and improved in the future.

We also faced pricing and time constraints when training our legal document summarization models. Fine-turning a large language model has a high computational cost which caused the models to take many hours to train. This was further emphasized by our training data containing long legal documents with an average of about 15000 words. Because of this, we trained our models using powerful cloud computing architecture which significantly reduced our training time. Using these resources did introduce a financial cost to training, which was expensive because we needed to operate high-performance hardware for long periods of time. This price restriction limited our training to the two successful runs that we incorporated into this paper. While we were able to see meaningful results from our models, it would be beneficial to train more models in the future and compare them.



(Figure 5: Relationship between BERT score and LLM-given score for legal document summaries)

Our project also faces limitations related to the data used to train our models. Our training dataset mostly consisted of legal documents from The Supreme Court of India, with some from the Supreme Court of the United Kingdom for universality. When applying our model for United States legal documents, the summarization may not be sufficient because they have different patterns. The best training results for legal document summarization of our training data may not be the best results for summarizing generic legal documents, and we do not have conclusive results for our model's performance on legal documents outside of these countries. Token

limitations present another significant challenge when working with long and complex legal documents. Legal texts often exceed the token capacity of most models, requiring the input to be segmented into smaller chunks. This segmentation can disrupt the model's ability to maintain coherence and contextual understanding across sections, leading to suboptimal results. Additionally, the token cap limits the model's ability to analyze multi-section documents holistically, often necessitating external summarization or context-handling strategies, which introduce additional complexity and potential inaccuracies. This also restricted our ability to train smaller, more bare-bones models which influenced our decision to use DistilBART for training and comparison.

Finally, we propose some further directions for this work. To further improve the model, more datasets of legal documents from other countries in English can be inserted to allow the model to recognize more tokens or patterns to summarize legal documents better. Also, we could apply some other useful techniques, such as token compression techniques and external memory modules that allow models to handle multi-section texts more effectively. For evaluation, developing holistic metrics that capture logical consistency, adherence to legal precedents, and practical applicability ensures models are tested in real-world scenarios, improving overall performance. Additionally, this is a semester-long project, and also because we are more focused on attempting different novel algorithms and testing their performance, we only distribute reasonably long enough time for each attempt. Spending more time on training the model will increase the overall performance of the model but it will take more time to improve the overall performance as we have spent long enough time to train it to converge.

7. CONCLUSION

In short, our project successfully develops a framework for using large language models to evaluate summaries which we applied to our own trained legal document summarization models. Through experimental analysis, our framework shows that we can achieve a semi-strong positive correlation between human judgments and LLM-generated evaluations when using GPT-4o to evaluate coherence, fluency, and consistency. Our legal document summarization models show mixed results, with some improvements over the baseline models and some areas for continued improvement. Our first model improves the accuracy of legal document summaries shown through considerable improvement in consistency, but shows some decreases in performance in relevance and fluency. Our second model improves upon the baseline model b, but loses some ability incorporating more legal information but loses some of its ability to writing quality English text which suggests that the model is overfitting the training data. Despite this, our evaluation underscores the importance of our proposed evaluation framework as it helps to uncover aspects of summarization model performance that go undiscovered using traditional scoring techniques such as BERT score. Both the evaluation framework and our legal document summarization models have areas for improvement, they are a useful step in understanding and optimizing text summarization using large language models.

8. REFERENCES

[1] Lin, C.-Y. (2004, July). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74–81. Retrieved from https://aclanthology.org/W04-1013

[2] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020, February 24). *BERTScore: Evaluating text generation with Bert*. arXiv.org. https://arxiv.org/abs/1904.09675

[3] Deroy, A., Ghosh, K., & Ghosh, S. (Year). *Applicability of Large Language Models and Generative Models for Legal Case Judgement Summarization*. Computer Science and Engineering, IIT Kharagpur, & Computational and Data Sciences, IISER Kolkata.

[4] Bhattacharya, P., Hiware, K., Rajgaria, S., Pochhi, N., Ghosh, K., Ghosh, S.: A comparative study of summarization algorithms applied to legal case judgments. In: Proc. European Conference on Information Retrieval (ECIR), pp. 413–428 (2019)

[5] Nigam, S.K., Deroy, A.: Fact-based court judgment prediction. arXiv preprint arXiv:2311.13350 (2023)

[6] Schaik, T. A. van, Pugh, B., Tempest A. van SchaikMicrosoft, R., & Brittany PughMicrosoft, R. (2024, July 11). *A field guide to Automatic Evaluation of LLM-generated summaries: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Conferences.

[7] Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). SummEval: Re-evaluating Summarization Evaluation. ArXiv:2007.12626 [Cs]. https://arxiv.org/abs/2007.12626