

# Precision-Driven Sentence Filtering for Long Text Summarization

Alex Mei, Anisha Kabir, John Judge, Rukmini Bapat | Advisors: Tony Sun, Professor William Wang

## Introduction

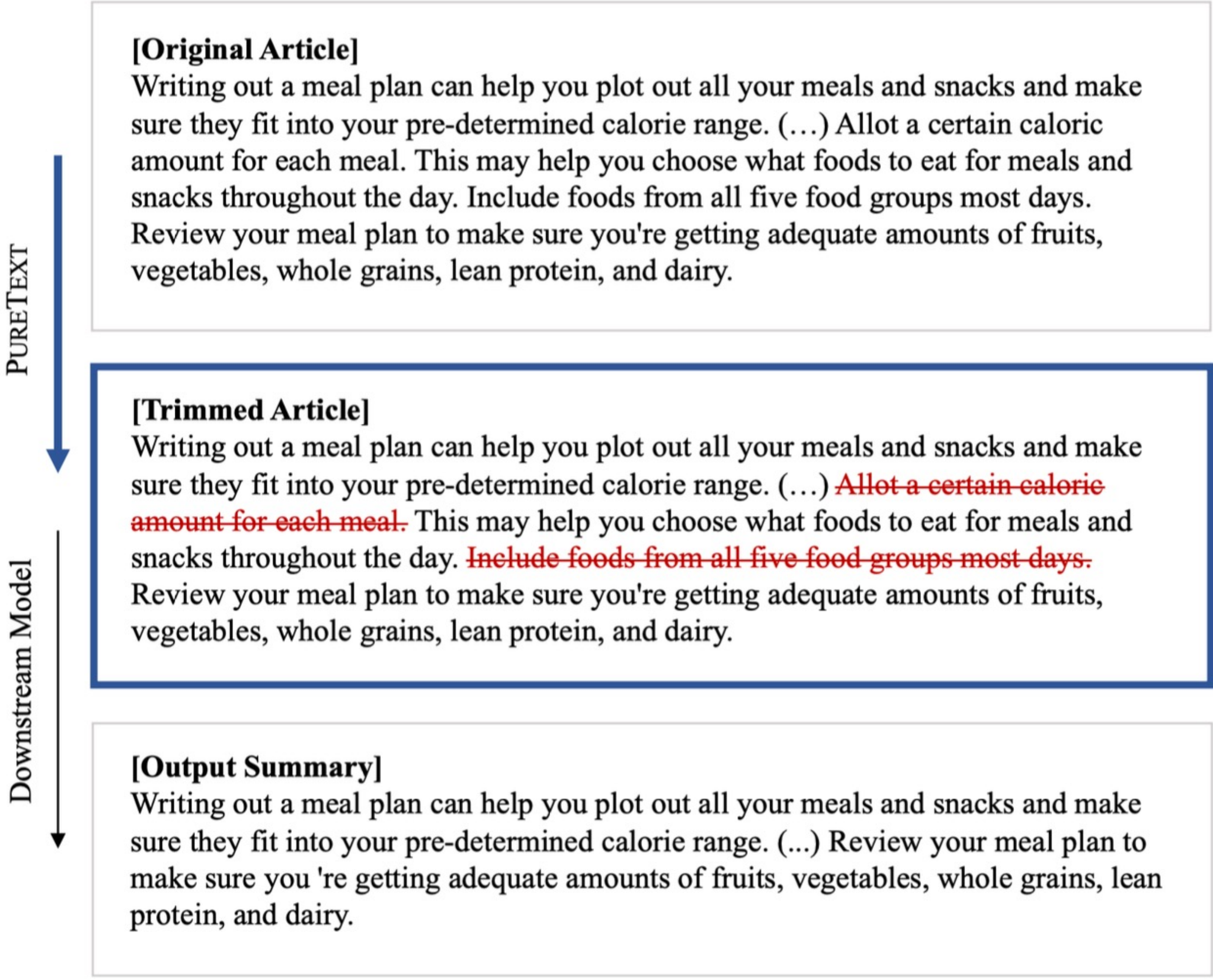
- Summarization models are limited by their maximum input length, posing a challenge to summarize longer texts comprehensively.
- Truncation may leave out critical parts of the text, leading to an incomplete summary.

## Research Goals

- To improve the performance and quality of summarization models on long texts.
- To identify a dataset- and model-agnostic approach to text filtering that fits within a summarization model's input limit.

## Methodology

- We introduce **PURETEXT**, a novel, lightweight framework for selecting high-quality sentences as a filtering step in long text summarization.
- We fine-tune a **BERT**-based model to classify sentences as either important or unimportant using a sentence's **ROUGE** score to generate its label.
- Using the 0-1 Knapsack algorithm we find the most important sentences up to the token limit.



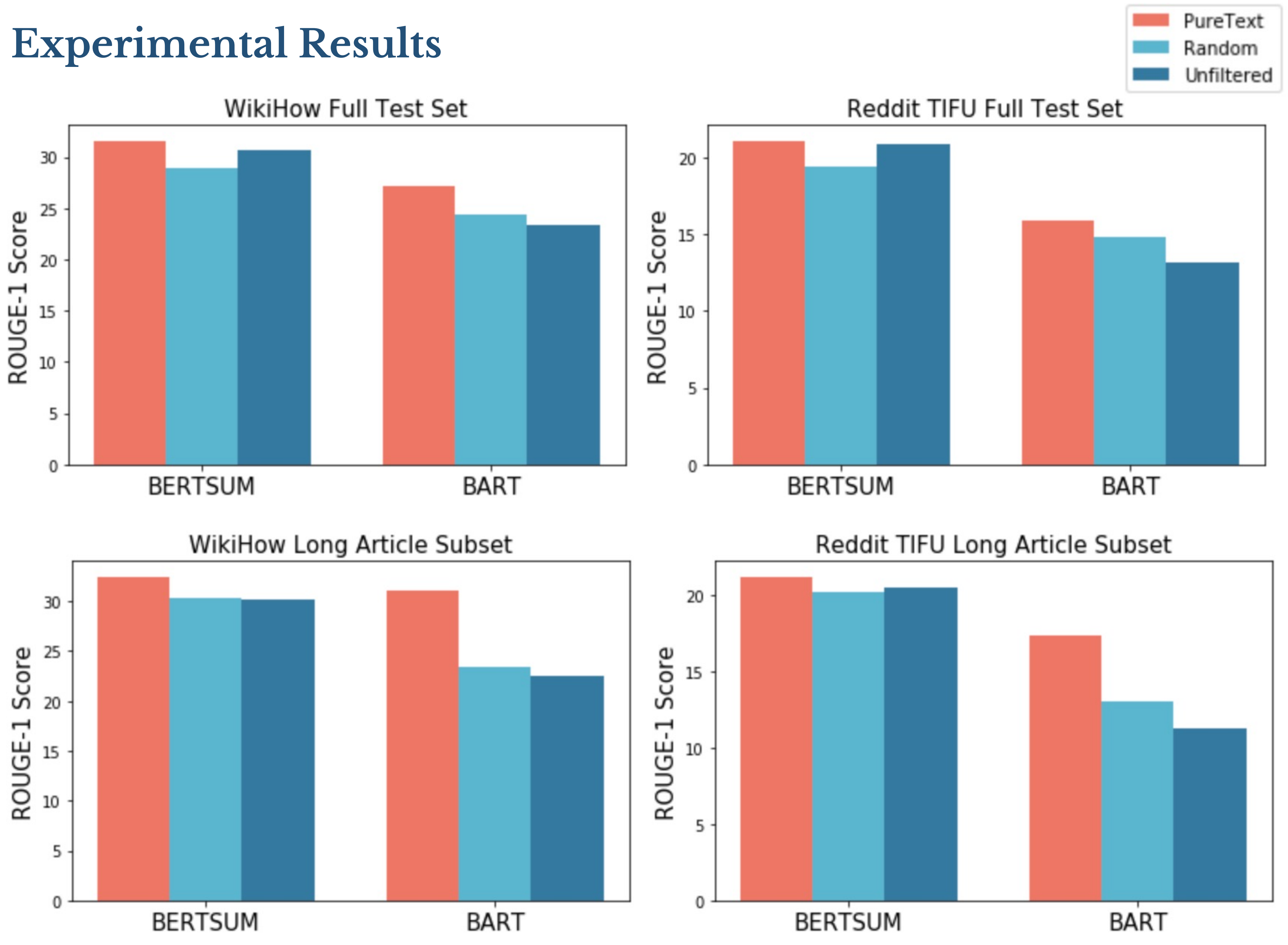
## Background

- Self-Supervision:** a form of machine learning that does not require human labeled data.
- BERT:** a deep neural model pre-trained to gain bidirectional context of unlabeled text.
- ROUGE:** a widely used recall-based summary evaluation metric; it reports the similarity between candidate and ideal summaries.

## Setup

- We test **PURETEXT** on non-journalistic datasets **WikiHow** and **Reddit TIFU**, and on downstream models **BERTSUM** and **BART**.
- We experiment on the **full** test dataset and the long article **subset** of each dataset.
- We test against baselines without **PURETEXT**: **unfiltered** and **random** dropping.

## Experimental Results



## Analysis

- PURETEXT** improves on the long text subset by a factor of 3 greater than the full dataset.
- These improvements provide statistically significant evidence ( $p < .05$ ) that **PURETEXT** improves long article summarization.
- PURETEXT** is particularly effective on long articles because truncation of these articles results in removing important sentences.
- PURETEXT** is most applicable to datasets like **WikiHow** and **Reddit**, where key sentences are evenly distributed.

## Conclusion

- We utilize a **BERT**-based model trained with self-supervised learning to distinguish high-quality sentences, which are then passed to a downstream summarization model.
- Our results show that **PURETEXT** can greatly improve upon downstream model baselines for multiple datasets and model. **PURETEXT** excels at long article summarization.
- We encourage future work to continue exploring the dataset and model-agnostic nature of such a sentence filtering approach.