**Summary Report: Performance of Pre-trained Text Summarization Models** horizontal line

This report provides an overview of the performance of three pre-trained text summarization models on news articles: BERT Summarizer, GPT-2, and XLNet. The models were evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric, which measures the quality of generated summaries against a reference summary. The average ROUGE-1, ROUGE-2, and ROUGE-L scores were calculated for each model using a dataset of 100 articles and their corresponding reference summaries. An average ROUGE F1-score of 0.4 is considered "good" for text summarization performance.

Results:

**BERT Summarizer:**

ROUGE-1: Recall = 0.3578, Precision = 0.1758, F1-score = 0.2323

ROUGE-2: Recall = 0.1158, Precision = 0.0517, F1-score = 0.0701

ROUGE-L: Recall = 0.3278, Precision = 0.1612, F1-score = 0.2129

**GPT-2:**

ROUGE-1: Recall = 0.3547, Precision = 0.1830, F1-score = 0.2369

ROUGE-2: Recall = 0.1070, Precision = 0.0523, F1-score = 0.0687

ROUGE-L: Recall = 0.3199, Precision = 0.1650, F1-score = 0.2136

**XLNet:**

ROUGE-1: Recall = 0.3509, Precision = 0.1755, F1-score = 0.2302

ROUGE-2: Recall = 0.1099, Precision = 0.0500, F1-score = 0.0674

ROUGE-L: Recall = 0.3158, Precision = 0.1579, F1-score = 0.2071

None of the models reached the "good" threshold of an average ROUGE F1-score of 0.4. The differences between the models were also relatively small, indicating that further fine-tuning using training data may lead to improved performance.