The Impact of Synthetic Data on Text Summarization Quality

This project explores the role of synthetic data in enhancing the quality of automatic text summarization. Synthetic datasets, often generated using large language models, provide a promising solution to address data scarcity and improve model generalization. However, challenges such as domain-specific applicability and evaluation complexities remain. The project will design and implement synthetic data generation techniques, evaluating their integration into summarization workflows and measuring the impact on model performance.

1 Introduction

Table 1 provides a comparative analysis of existing solutions for automatic text summarization with a focus on the role of synthetic data.

Table 1: Comparative analysis of basic solution

Solution	Strengths	Weakness
Synthetic data generation with LLMs for text classification [1]	LLMs effectively generate synthetic datasets, especially in few-shot settings. Promising results for low-subjectivity tasks demonstrate its utility in specific domains.	Synthetic data struggles with high-subjectivity tasks and lacks diversity compared to real-world data. Dependence on real examples and high computational costs limit broader adoption.
Using SFT and RLHF for summarization in Russian [2]	Synthetic datasets improve summarization quality and address data scarcity in low-resource languages. Shows that synthetic data paired with RLHF, can align models more closely with human preferences.	High-quality synthetic data requires significant resources and struggles with real-world variability. Overreliance on synthetic data can hinder generalization and practical scalability.

References

[1] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. *ArXiv*, abs/2310.07849, 2023.

[2] Albina Akhmetgareeva, Alexander Abramov, Ilia Kuleshov, Vlad Leschuk, and Alena Fenogenova. Towards russian summarization: can architecture solve data limitations problems?, 2024.

[3] Connor Shorten, Taghi M. Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of Big Data*, 8(1), July 2021.

[4] Yingzhou Lu, Huazheng Wang, and Wenqi Wei. Machine learning for synthetic data generation: a review. *CoRR*, abs/2302.04062, 2023.

[5] Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. Automatic text summarization methods: A comprehensive review. *CoRR*, abs/2204.01849, 2022.

[6] Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *CoRR*, abs/2403.02901, 2024.

[7] Jay Ahn and Foaad Khosmood. Evaluation of automatic text summarization using synthetic facts. *CoRR*, abs/2204.04869, 2022.

[8] Mandeep Goyal and Qusay H. Mahmoud. A systematic review of synthetic data generation techniques using generative ai. *Electronics*, 13(17):3509, September 2024.

[9] Meng Cao. A survey on neural abstractive summarization methods and factual consistency of summarization, 2022.