

UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

**Abstractive summarization of news in
Ukrainian**

Author:
Yuliia VOZNIAK

Supervisor:
Oleksii IGNATENKO

*A thesis submitted in fulfillment of the requirements
for the degree of Master of Science*

in the

Department of Computer Sciences
Faculty of Applied Sciences



Lviv 2024

Declaration of Authorship

I, Yuliia VOZNIAK, declare that this thesis titled, “Abstractive summarization of news in Ukrainian” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Abstractive summarization of news in Ukrainian

by Yuliia VOZNIAK

Abstract

Significant advancements in large language models in recent years have brought rapid improvements in all areas of text processing, including text summarization. However, in the case of low-resource languages like Ukrainian, this improvement is underresearched. In this thesis we take a snapshot of the current state of abstractive text summarization of Ukrainian news. To achieve this, we review the datasets available for the Ukrainian language, highlight existing problems and propose a possible improvement by introducing a custom news summarization dataset created by matching summaries from Telegram channels with their respective articles on the news media websites. We also experiment with text summarization via LLM Llama-3 and analyze the results both quantitatively and qualitatively.

Contents

Declaration of Authorship	ii
Abstract	iii
1 Introduction	1
1.1 Thesis structure	1
2 Related work	3
2.1 Factual consistency	3
2.2 Metrics	4
2.2.1 Abstractive summarization in Ukrainian	6
3 Data exploration	7
3.1 Existing datasets overview	7
3.1.1 Problems with existing data	7
3.2 Custom dataset creation	8
3.2.1 Data gathering	8
3.2.2 Preliminary data cleaning	9
3.3 Data quality evaluation and selection	9
3.3.1 First round	10
Descriptive statistics	10
Extractive fragments-based	10
Filtering criteria	11
3.3.2 Second round	12
BERTScore	12
Updated filtering criteria	14
4 Approach	15
4.1 Fine-tuning	15
4.1.1 LORA	15
4.1.2 QLORA	16
4.2 Experimental setup	16
LLMs selection	16
Text generation pipelines	16
Hardware setup	16
5 Experiments	17
5.1 Zero-shot	17
Prompt selection	17
Analysis of results	17
5.2 Experiments	18
5.2.1 Llama finetuning	18
Base set of parameters	18

Experiment 1	18
Experiment 2	19
Experiment 2.2	19
Experiment 3	19
Experiment 4	19
Experiment 5	19
Prompt-answer format conformity	19
5.2.2 Mistral finetuning	19
5.2.3 Inference	20
5.3 Evaluation	20
5.4 Problems	21
5.5 Interesting findings and conclusions	24
6 Conclusions and discussion	26
Bibliography	27

List of Figures

3.1	Example from XL-Sum dataset. The summary (leading paragraph) contains information absent from remaining article. All words that contain "oil" root are highlighted in in bold and are respectively: oil pipeline, oil supply and oil producing. These words are highly specific terms and it is clear that the first paragraph (summary) contains information, absent from the later part. This can lead to the conclusion that training on such data will only foster hallucinations.	8
3.2	Example from the created dataset	9
3.3	Pseudocode describing the process of calculating the set of extractive fragments $F(A, S)$ in summary S , which were extracted from article A . (Image from [Grusky, Naaman, and Artzi, 2018])	11
3.4	BERTScore for a text-summary pair from [Hasan et al., 2021], which was mentioned in 3.1. Score is much higher for the English language, probably because in Ukrainian words with "oil" root are different separate words, while in English they are two-word combination with one word being "oil". It was stated previously, that in this case summary does not entail the reference text. However, they are thematically similar.	12
3.5	A summary-article pair from our custom dataset. The English version shows much higher score than the Ukrainian one.	13
3.6	Summaries in summary-article pairs are swapped.	13
3.7	Summaries in summary-article pairs are swapped.	13
3.8	Distributions of these characteristics of the unfiltered dataset. Black lines represent the threshold of the second filtering (either lower or upper). These plots also do not include outliers.	14
5.1	The prompt-answer format that was used for training and inference.	17
5.2	Example of a zero-shot text summarization by Llama3 8B. The second row includes the prompts. It is clear that for both cases model generated possible continuation of the text with zero regards to attempts to ask it to do something specific.	18
5.3	An example showing a summary generated by a model with less prominent repetition problem and the one, where this is a major issue.	21
5.4	Generated summary(left) and input article (right). Settlement names are highlighted in the reference text. This case shows how the model can hallucinate but produces seemingly accurate statements as the locations mentioned in the summary all are valid places, located near the frontline in Donetsk oblast. The problem is only that they were not mentioned in the article.	22

5.5	Three generated summaries from different models on the left and reference article on the right. Clear examples of hallucinations are highlighted. The messages delivered in the summaries are only partially related to the actual meaning of the article. However, even with little training, the models were able to produce results plausible enough to be confusing.	22
5.6	These summaries (left) present several common problems: repetition and clutter. It also includes transformation of an emotional quote into a statement without corresponding attribution to the author.	23
5.7	Example of a good summary that is simultaneously complete, well-formatted and factually correct.	24
5.8	Example of two summaries that independently generate similar factually inconsistent information.	24
5.9	Example showcasing ability to make correct abbreviations without losing context.	25

List of Tables

3.1	Descriptive statistics	10
3.2	Extractive fragments-based statistics	11
5.1	Average metrics for different experiments. R-1, R-2, R-3 denote ROUGE-1, ROUGE-2 and ROUGE-L respectively.	21
5.2	The discrepancy between rouge-family scores stated in Hasan et al., 2021 and Galeshchuk, 2023 is quite puzzling.	21

List of Abbreviations

LLM	Large Language Model
NLI	Natural Language Inference
NLI	Natural Language Processing
RNN	Recurrent Neural Network

Chapter 1

Introduction

Abstractive text summarization is the NLP task of creating a summary that conveys the main points of the original larger text in several sentences. The abstractiveness of a summary means that it is not just an extraction of a part of the original text verbatim and paraphrasing and condensing were used instead. It is contrasted with an extractive summarization approach, which involves selecting the most important sentences from the original text and assembling them into a summary.

Due to the abovementioned paraphrasing, condensing and variability of language, the task of Abstractive Summarization has an inherent challenge at its core: there are many ways of creating a correct summary and different people will have different versions of a «gold-standard summary» for each particular text. This fact makes evaluating summarization results difficult as even human evaluation, while being time-and-labor-consuming, can simultaneously be very subjective and biased. The problem is a bit narrower in the case of news summarization as news articles typically focus on a self-contained event, making it easier to distinguish the main points and formulate them into a summary.

Recent years have seen an impressive breakthrough in the performance of generative AI overall and Large Language Models in particular. They can confidently produce a summary, that at least at first sight is correct and credible. However, the generative nature of the task inevitably leads to the possibility of hallucinations i.e. grammatically and lexically correct but factually wrong output.

The progress in the sphere of LLMs is happening so rapidly, that there is few (publicly) available existing relevant research, especially for a low-researched language like Ukrainian.

To contribute to mitigating this gap of research, in this thesis we focused on the current state of abstractive summarization of news in Ukrainian: examined the existing datasets and solutions, highlighted possible problems, suggested another approach for dataset creation, experimented with recently released Llama-3 LLM and analyzed generation results.

1.1 Thesis structure

1. **Related work** Chapter 2 contains a brief exploration of the different periods of abstractive text summarization research including an overview of some of the common metrics for summarization. Special attention is being paid to the research of quality improvement, factchecking and minimizing large language models hallucinations, which became a prominent issue accompanying the advancements in recent years.
2. **Data overview** Chapter 3 provides an overview of existing datasets, that include data for the Ukrainian language and pertinent analysis and criticism.

It also contains a detailed description of the custom dataset creation process, including data gathering, exploratory data analysis and several stages of determining the filtering criteria applied for data quality improvement.

3. **Approach description** Chapter 4 elaborates on the specific details of the experimentation approach. It contains descriptions of the fine-tuning methods and their parameters, selected LLMs and hardware setup.
4. **Experiments** Chapter 5 provides details of the experimental setups, training process, results and their analysis. It also contains description of evaluation metrics and results achieved by using fine-tuned models. Those results are compared with existing scores from the previous work on summarization in Ukrianinan.

Chapter 2

Related work

In this section we briefly address important related work with focus on the advancement of approaches to abstractive summarization, overview of common evaluation metrics for summarization and research on reducing hallucinations in summarization outputs.

In the early era of summarization, researchers focused on selecting important sentences taken verbatim from the input and composing a summary from these extractions [Nenkova, 2011]. Introduction of RNNs and sequence-to-sequence models made it possible to combine extractive and abstractive approaches Rush, Chopra, and Weston, 2015, [Nallapati et al., 2016, See, Liu, and Manning, 2017] and use some paraphrasing. Despite this progress, at that time good performance was only achieved on short input and output sequences. When used with longer documents and summaries, these models often would output repetitive and incoherent phrases [Paulus, Xiong, and Socher, 2017]. To facilitate further research, datasets for abstractive summarization were created [Grusky, Naaman, and Artzi, 2018, Narayan, Cohen, and Lapata, 2018a inter alia].

The introduction of transformers [Vaswani et al., 2017] marked a new phase in text summarization with a novel ability to generate truly abstractive summaries. New state-of-art solutions emerged [Cohan et al., 2018, Liu and Lapata, 2019 inter alia]. Advancements in models' abilities fostered interest in exploring multilingual summarization [Scialom et al., 2020, Fabbri et al., 2019]. Several large multilingual datasets emerged of varying quality [Fabbri et al., 2019, Varab and Schluter, 2021, Hasan et al., 2021, Ladhak et al., 2020]. Great improvements in summarization were introduced by [Zhang et al., 2020, Raffel et al., 2023]. After the introduction of GPT-3 and the later developments, many works focused on using it as a summarizer or factuality evaluator [Goyal, Li, and Durrett, 2023, Luo, Xie, and Ananiadou, 2023, Zhang, Liu, and Zhang, 2023 inter alia].

2.1 Factual consistency

The improved abilities to create coherent and fully fluent texts exacerbated the inherent problem of text generation overall: large language models are prone to hallucinations i.e. producing plausible but factually incorrect information. This underlying issue makes real-life usage of text generation impractical, as there is a probability of serious and meaningful factual inconsistencies. In order to avoid such thing, many approaches and models were suggested to enable the creation of models with better factual consistency.

One of the first such examples was FactCC - a weakly supervised model for identifying conflicts between source documents and a generated summary [Kryściński et

al., 2019]. Authors fine-tuned a base BERT model on a custom dataset, which combined manually created examples with synthetic data, made by applying transformations (paraphrasing, sentence negation, entity, number and pronouns swapping, adding noise) to selected statements from the summary, thus making it factually inconsistent. The model was trained as a binary classifier to judge whether the summary is factually consistent with the larger text.

The publication "On Faithfulness and Factuality in Abstractive Summarization" [Maynez et al., 2020] delved into differences between internal, external and factual hallucinations and suggested that entailment measures are beneficial for improvement of factual consistency. To this end, they trained an entailment classifier based on BERT encoders, using a sentence-sentence NLI dataset as training data. Their model predicted the probability of the summary entailing the reference text and the scores were then compared with human evaluations of factuality and faithfulness. Reducing named entity hallucinations was in the focus of the research by [Akani et al., 2023]. The suggested approach involves detecting named entities and understanding whether they can be linked to other named entities within the reference article with the main idea being that Out-Of-Document entities are more likely to be hallucinations. The logical relationship between the reference and the summary was researched by [Aharoni et al., 2023] where they tested entailment between summary-article pairs of the XL-SUM dataset [Hasan et al., 2021] and got predominantly results of low entailment between them. The authors also suggested a novel multilingual NLI model to improve factual consistency in abstractive summarization.

Another aspect of consistency between reference article and a summary was tested in the publication "Abstractive Summarizers Become Emotional on News Summarization" [Ahuir et al., 2024]. The authors analyzed whether emotionally charged words will be transferred into the summary or it will be more neutral. They also examined the distribution and frequency of different emotional sentiments in the news articles and their respective summaries.

2.2 Metrics

Metrics, that are used to evaluate summarization quality were researched long before developments of technology made it possible to generate abstractive summaries automatically. First text generation metrics were calculated on basic features like single words, n-grams, or other text statistics. Later, it became feasible to train separate models for summarization evaluation or to use word embeddings like BERT [Devlin et al., 2018]. Notwithstanding this novel advancements, the classic metrics are often used for comparison even in the most recent papers with one of the main contributing factors being that these new metrics often were relevant only for the English language.

BLEU score [Papineni et al., 2001], formulated as a method to assess translation quality automatically. BLEU (bilingual evaluation understudy) is calculated by comparing a target string with a set of reference strings from the corpus and measuring their n-gram overlap. The overall score is a weighted geometric mean of all n-gram precisions, thus making it favor candidates that have good scores based on multiple n-grams. This baseline metric was also enhanced to be more robust by multiplying the result by a brevity penalty, which is 1 if the generated text is longer than the reference text and otherwise $brevity\ penalty = e^{(1 - length\ generated / length\ reference)}$. BLEU score was long used as the main metric for machine translation and was often utilized for text summarization.

$$p_n = \frac{\sum_a^b \sum_{s \in G_n(\hat{y}^{(i)})} \min(C(s, \hat{y}^{(i)}), \max_{y \in S_i} C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y}^{(i)})} \quad (2.1)$$

In 2.1 $C(s,y)$ denotes substring count – the number of appearances of s as a substring of y , \hat{y} is a candidate text, y is a reference text, and G_n is a set of its unique n -grams.

$$BLEU = \text{Brevity penalty} \times \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (2.2)$$

ROUGE [Lin and Hovy, 2003] (Recall-Oriented Understudy for Gisting Evaluation) is a family of metrics, that was introduced in 2003 and is used for evaluation of text summarization and machine translation. It is calculated as a recall between input summary and a reference summary or a set of reference summaries. ROUGE-N (common variants: ROUGE-1, ROUGE-2) measures the recall of unigrams and bigrams respectively, while ROUGE-L considers longest common subsequences.

$$ROUGE - N = \frac{\sum_{S \in \{\text{Reference summaries}\}} \sum_{gram_n} \text{count}_{\text{match}}(gram_n)}{\sum_{S \in \{\text{Reference summaries}\}} \sum_{gram_n} \text{count}(gram_n)} \quad (2.3)$$

With n standing for the length of the n -gram, $gram_n$, and $\text{count}_{\text{match}}(gram_n)$ is the maximum number of n -grams co-occurring in a candidate summary and a set of reference summaries.

$$p_{lcs} = \frac{(1 + \beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \quad (2.4)$$

$$R_{lcs} = \frac{LCS(X, Y)}{m} \quad (2.5)$$

$$P_{lcs} = \frac{LCS(X, Y)}{n} \quad (2.6)$$

METEOR [Lavie and Agarwal, 2007] (Metric for Evaluation of Translation with Explicit ORdering) was proposed in 2005. Calculation includes a harmonic mean of unigram precision, recall, F-score, bigram overlap, and exact word matches.

$$P = \frac{m}{w_t} \quad (2.7)$$

$$R = \frac{m}{w_r} \quad (2.8)$$

$$F_{\text{mean}} = \frac{10PR}{R + 9P} \quad (2.9)$$

m : Number of unigrams in the candidate translation also found in reference, w_t : Number of unigrams in candidate translation, w_r : Number of unigrams in reference translation

$$p = 0.5 \times \left(\frac{c}{u_m}\right)^3 \quad (2.10)$$

C : Number of chunks in candidate, U_m : Unigrams in candidate

$$METEOR = F_{\text{mean}}(1 - p) \quad (2.11)$$

BERTScore was introduced in 2019 as an automatic evaluation metric for text generation [Zhang et al., 2019]. It utilizes pretrained BERT [Devlin et al., 2018] embeddings and computes cosine similarity between all tokens in target and input texts.

These and other metrics (a total of 24) were analyzed in [Fabbri et al., 2020] by their correlation with coherence, consistency, fluency and relevance of the summary, with those reference metrics being determined by human evaluation.

2.2.1 Abstractive summarization in Ukrainian

While there is an extensive research on abstractive summarization for the English language, there is, to my best knowledge, only one paper for Ukrainian Galeshchuk, 2023. Galeshchuk S. fine-tuned a multilingual T5 model [Raffel et al., 2023] on the custom data. The dataset was created by pairing first paragraphs of articles on Hromadske news website with the remainder of the respective article. To our best knowledge the dataset is not publicly available so can not be used for comparison or integrated into the training. The author compared the summaries generated by the model trained on their data and the model trained on XL-SUM dataset. For evaluation ROUGE-family metrics were used and the values will be included for comparison later in our paper.

Chapter 3

Data exploration

In this chapter, we provide details of the custom dataset together with the description of the creation and filtering process. In 3.1 summarization datasets that support Ukrainian are overviewed and several existing problems are highlighted. In section 3.2 the process of custom dataset creation is described.

3.1 Existing datasets overview

While there are many publicly available summarization datasets of variable quality for English language, the situation is very different for low-resource languages like Ukrainian. To my best knowledge, there are two public summarization datasets, that include data in Ukrainian language: MassiveSumm [Varab and Schluter, 2021] and XL-SUM [Hasan et al., 2021]. Creation of another dataset was described in [Galeshchuk, 2023], however it is not publicly available.

XL-SUM [Hasan et al., 2021] was created to be a multilingual counterpart to the English-only X-SUM dataset [Narayan, Cohen, and Lapata, 2018b]. In contrast, XL-SUM comprises data for 44 different languages, including 57 000+ article-summary pairs for Ukrainian. The data was gathered from different localization versions of the BBC website¹.

Massivesumm [Varab and Schluter, 2021] dataset was created based on the assumption, that html meta "description" property should describe the pertinent content and be a valid summary of the articles.

Ukrainian summarization dataset by [Galeshchuk, 2023] was created from articles by Hromadske in the way, similar to XL-SUM [Hasan et al., 2021] way.

3.1.1 Problems with existing data

The majority of summarization datasets (regardless of the target language) were created by extracting the leading paragraph of the article [Hasan et al., 2021, Narayan, Cohen, and Lapata, 2018b inter alia] or by using the "description" meta attribute of an html page [Varab and Schluter, 2021]. However such approaches can foster hallucinations.

In particular, the "leading paragraph" approach has an important drawback: while beginning of an article was professionally written and often contains the gist of the following article, this does not necessarily mean that it can be used separately. For example, the first paragraph may provide some necessary contextual details and, assuming that the reader already got this information previously, do not repeat them later.

¹<https://www.bbc.com/>

"summary": "Горить **нафтопровід** на півдні Іраку. **Нафтопостачання** в тій частині країни припинене.",

"text": "За повідомленнями агенцій, у **нафтовидобувному** місті Кіркук від рук повстанців загинули один громадянин Південної Африки, один - Нової Зеландії та один іракець. За повідомленнями з Багдада, щонайменше шестеро людей, у тому числі четверо іноземців, зазнали поранень внаслідок вибуху. Вибух стався біля готелю в центрі міста, де переважно мешкають іноземні працівники. Вибух стався кількома годинами по тому, як принаймні семеро людей загинули і ще кілька десятків зазнали поранень внаслідок двох вибухів на одному зі столичних ринків. У кількох містах Іраку тривали сутички між силами коаліції та прихильниками шіїтського клерика Моктади аль-Садра. Американське командування повідомило про загибель понад 20 іракців, у тому числі кількох службовців поліції."

FIGURE 3.1: Example from XL-Sum dataset. The summary (leading paragraph) contains information absent from remaining article. All words that contain "oil" root are highlighted in bold and are respectively: oil pipeline, oil supply and oil producing. These words are highly specific terms and it is clear that the first paragraph (summary) contains information, absent from the later part. This can lead to the conclusion that training on such data will only foster hallucinations.

This problem was extensively researched in the Multilingual Summarization with Factual Consistency Evaluation work [Aharoni et al., 2023]. They used their multilingual NLI model on the XLSUM training dataset, calculating the number of examples, where the summary was, in fact, entailed by the input. For the Ukrainian language their calculations showed that only 38.2 % of summaries were factually consistent. While some errors may be attributed to the NLI model itself or to the fact, that Ukrainian language was not represented in the NLI dataset, on which the model in question was trained, the situation is not drastically better for other languages. Furthermore, selecting a random article-summary pair can prove the following statistic quite well as shown in 3.1.

Using "description" meta attribute is not optimal either: usually it either is identical to the leading paragraph, or contains text, which was optimised for searching purposes, thus making it unsuitable for summarization dataset.

To avoid this negative effect, we decided to get data from Telegram channels of Ukrainian news agencies, where summaries are created to be self-contained separate entities and thus can be used apart from the main article.

3.2 Custom dataset creation

3.2.1 Data gathering

In order to choose specific channels to get summaries from, the following selection criteria was applied:

- a) The telegram channel should be a representation of an established large news media resource and not an independent opinion-based blog.
- b) The telegram channel should contain both a message with short summary of a piece of news and a url, that leads to the full article on the respective website.
- c) The summaries in question should be relatively short and unique i.e. they should not repeat the leading paragraph of the article verbatim.

d) The summaries should be written in a rather "neutral style", without emotional calls to action, questions or other misleading descriptions added for shock value or higher engagement with content.

FIGURE 3.2: Example from the created dataset

```
"text": "18 травня о 22:00 почався перший півфінал Пісенного конкурсу Євробачення-2021 – гурт Go_A виступили під номером 15. Глядачі побачили виступи 16 учасників і переможця Євробачення-2019 Дункана Лоуренса. Також співачка Давіна Мішель і актриса Текла Ройтен показали перфоманс на тему сили води: за допомогою доповненої реальності стіни арени Аhoу вкривуться водою. До фіналу також гарантовано увійшли країна-господарка і \"велика п'ятірка\": Зазначимо, за рейтингом букмекерів, гурт Go_A потрапив у п'ятірку потенційних фіналістів Євробачення-2021. Дивіться Пісенний конкурс Євробачення-2021 — другий півфінал 20 травня і фінал — 22 травня о 22:00 на Суспільному та вболівайте за учасника від України гурт Go_A, який виступить з піснею SHUM під номером 15. Нагадаємо, в Україні пряму трансляцію Пісенного конкурсу Євробачення показують суспільний телеканал UA: ПЕРШИЙ та комерційний телеканал СТБ.",
"summary": "Україна вийшла до фіналу Євробачення — гурт Go_A потрапив до першої десятки фіналістів."}
```

After deliberation, the following channels were selected:

Name	telegram link	website	num. posts
Suspilne.News	t.me/s/suspilnenews	https://suspilne.media	30 000+
Hromadske	t.me/s/hromadske_ua	https://hromadske.ua	43 000+
Espreso TV	t.me/espresotb	https://espreso.tv	62 000+

3.2.2 Preliminary data cleaning

To avoid including irrelevant data, preliminary data cleaning process was performed by removing all posts that

- contain links to multiple articles
- do not contain any links
- contain links to articles on other websites
- contain links that lead to resources other than single article.

After this steps, the remaining data consisted of about 71 000 posts total with the following breakdown by source: Suspilne.News - 15 000, Hromadske - 22 000, EspresoTV - 33 000.

3.3 Data quality evaluation and selection

While preliminary data cleaning removed entries that were obviously irrelevant, this is not sufficient for creation of a summarization dataset that will actually improve model's summarization skills. However, there is no absolute metric of a "good summary" because the task of assessing summary quality is not only quite subjective but also can have multiple correct answers. Therefore, to filter out worse summaries, a combination of several criteria must be applied. To address this issue, many possible approaches and metrics were introduced in order to assess the quality of a reference summary. They can be broadly divided into several categories: descriptive statistics, metrics based on n-grams, based on pretrained embeddings or complex metrics, that can be calculated by separately trained models. Unsurprisingly, the more complex metrics can provide better assessments of summary quality but they almost always support only English language.

TABLE 3.1: Descriptive statistics

	# docs	avg article length		avg summary length	
		tokens	sentences	tokens	sentences
Hromadske	22 313	709.57	41.39	36.9	2.06
Suspilne	15 360	369.11	22.22	37.04	2.05
Espresso	33 960	216.58	11.09	36.56	2.03
Total	71 633	402.85	22.91	36.79	2.04

3.3.1 First round

Descriptive statistics

Descriptive statistics comprise simpler methods of extracting features from the text. Due to the universality and simplicity, they can be performed for any language. While it does not consider semantic meanings and paraphrasing, it can be valuable in selecting better quality summary-article pairs, namely texts that are too short or too long can be dropped. For our particular case, the breakdown of average article and summary lengths are shown in 3.1. For tokenization and sentence splitting a library "tokenize-uk"² was used.

Extractive fragments-based

Another approach, that is common with summarization evaluation, is n-gram overlapping and its derivatives. It was introduced in the paper related to a NEWSROOM dataset [Grusky, Naaman, and Artzi, 2018], which included several methods of assessing reference summary quality based on overlapping parts of the summary and the full article – extractive fragments. Those metrics were namely extractive fragment density, extractive fragment coverage and compression. One of their intended purposes was to quantify the difference between extractive and abstractive summaries and sort them into different categories. In our case, this can be used to filter out summaries, that include unaltered sentences from the reference article. To this end, another useful metric was proposed in "Intrinsic Evaluation of Summarization Datasets" [Bommasani and Cardie, 2020], which generalises extractive fragments coverage into less obscure "abstractivity" metric.

Extractive Fragment Coverage:

$$Coverage(A, S) = \frac{1}{|S|} \sum_{f \in F(A, S)} |f| \quad (3.1)$$

Extractive Fragment Density: The average length of the extractive fragment pertaining to each word in the summary.

$$Density(A, S) = \frac{1}{|S|} \sum_{f \in F(A, S)} |f|^2 \quad (3.2)$$

²<https://github.com/lang-uk/tokenize-uk>

```

function  $\mathcal{F}(A, S)$ 
   $\mathcal{F} \leftarrow \emptyset, \langle i, j \rangle \leftarrow \langle 1, 1 \rangle$ 
  while  $i \leq |S|$  do
     $f \leftarrow \langle \rangle$ 
    while  $j \leq |A|$  do
      if  $s_i = a_j$  then
         $\langle i', j' \rangle \leftarrow \langle i, j \rangle$ 
        while  $s_{i'} = a_{j'}$  do
           $\langle i', j' \rangle \leftarrow \langle i' + 1, j' + 1 \rangle$ 
        if  $|f| < (i' - i - 1)$  then
           $f \leftarrow \langle s_i \cdots s_{i'-1} \rangle$ 
         $j \leftarrow j'$ 
      else
         $j \leftarrow j + 1$ 
     $\langle i, j \rangle \leftarrow \langle i + \max\{|f|, 1\}, 1 \rangle$ 
     $\mathcal{F} \leftarrow \mathcal{F} \cup \{f\}$ 
  return  $\mathcal{F}$ 

```

FIGURE 3.3: Pseudocode describing the process of calculating the set of extractive fragments $F(A, S)$ in summary S , which were extracted from article A . (Image from [Grusky, Naaman, and Artzi, 2018])

TABLE 3.2: Extractive fragments-based statistics

source	number docs	average		
		abstractivity	density	compression
Hromadske	22 313	0.191	6.151	24.815
Suspilne	15 360	0.321	6.151	11.635
Espreso	33 960	0.2	13.650	8.842
Total	71 633	0.223	9.132	14.416

Compression: The token ratio between the article and summary.

$$\text{Compression}(A, S) = \frac{|A|}{|S|} \quad (3.3)$$

Abstractivity: Introduced in "Intrinsic Evaluation of Summarization Datasets" [Bommasani and Cardie, 2020], abstractivity generalizes extractive fragment coverage [Grusky, Naaman, and Artzi, 2018] to represent abstractivity of the summary.

$$\text{Abstractivity}(A, S) = 1 - \frac{1}{|S|} \sum_{f \in F(A, S)} |f| \quad (3.4)$$

Filtering criteria

For the purpose of improving the quality of data and filtering out outliers, the following criteria were applied:

- compression < 25

- abstractivity > 0.07
- tokens summary < 89
- 100 < tokens article < 571
- sentences summary < 5
- 1 < sentences article < 32

3.3.2 Second round

The results of the first experiment 5.2.1 suggested that further data filtering would be beneficial. In this round we applied filtering criteria based on the BERTScore, extractive fragments density and increased threshold for minimal abstractivity.

BERTScore

BERTScore is an automatic evaluation metric for text generation. Analogously to common metrics, BERTSCORE computes a similarity score for each token in the candidate sentence with each token in the reference sentence. However, instead of exact matches, token similarity is computed using contextual embeddings. [Zhang et al., 2019]

For calculation of BERTScore for texts in Ukrainian, a version of Ukrainian Roberta "youscan/ukr-roberta-base"³ was used and for calculation of BERTScore for texts in English, transformers' "google-bert/bert-base-uncased"⁴ model was used.

Several examples were analyzed in order to get some overview of how this score reflects different changes in text pairs.

texts in Ukrainian	BERTScore	texts in English	BERTScore
<p>За повідомленнями агентцій, у нафтовидобувному місті Кіркук від рук повстанців загинули один громадянин Південної Африки, один - Нової Зеландії та один іракець. За повідомленнями з Багдада, щонайменше шестеро людей, у тому числі четверо іноземців, зазнали поранень внаслідок вибуху. Вибух стався біля готелю в центрі міста, де переважно мешкають іноземні працівники. Вибух стався кількома годинами по тому, як прийаймі семеро людей загинули і ще кілька десятків зазнали поранень внаслідок двох вибухів на одному зі столичних ринків. У кількох містах Іраку тривали сутички між силами коаліції та прихильниками шіїтського клерика Моктади аль-Садра. Американське командування повідомило про загибель понад 20 іракців, у тому числі кількох службовців поліції.</p> <p>Горить нафтопровід на півдні Іраку. Нафтопостачання в тій частині країни припинене.</p>	0.5282	<p>According to news agencies, in the oil-producing city of Kirkuk, one citizen of South Africa, one citizen of New Zealand and one Iraqi died at the hands of the rebels. According to reports from Baghdad, at least six people, including four foreigners, were injured as a result of the explosion. The explosion occurred near a hotel in the city center, where mostly foreign workers live. The blast came hours after at least seven people were killed and dozens more injured in two explosions in one of the capital's markets. Clashes continued in several Iraqi cities between coalition forces and supporters of Shiite cleric Moqtada al-Sadr. The American command reported the death of more than 20 Iraqis, including several police officers.</p> <p>An oil pipeline is on fire in southern Iraq. Oil supply in that part of the country has been stopped.</p>	0.8095

FIGURE 3.4: BERTScore for a text-summary pair from [Hasan et al., 2021], which was mentioned in 3.1. Score is much higher for the English language, probably because in Ukrainian words with "oil" root are different separate words, while in English they are two-word combination with one word being "oil". It was stated previously, that in this case summary does not entail the reference text. However, they are thematically similar.

³<https://huggingface.co/youscan/ukr-roberta-base>

⁴<https://huggingface.co/google-bert/bert-base-uncased>

texts in Ukrainian	BERTScore	texts in English	BERTScore
Через загострення ситуації на Близькому Сході з сектора Газа виявили бажання вийхати 120 громадян України. Наразі посольства в Ізраїлі та Єгипті шукають способи їх репатріації. Про це Суспільному повідомив речник МЗС Олег Ніколенко. "Міністерство закордонних справ підтримує зв'язок із українцями, які виявили бажання залишити Сектор Газа. Зараз їх 120 осіб", — поінформував Ніколенко. За його словами, наразі посольства в Ізраїлі та Єгипті, а також представництво при Палестинській національній адміністрації відпрацьовують можливі маршрути репатріації. Речник додав, що МЗС тримає цю ситуацію на особливому контролі.	0.7267	Due to the aggravation of the situation in the Middle East, 120 citizens of Ukraine expressed their desire to leave the Gaza Strip. Currently, the embassies in Israel and Egypt are looking for ways to repatriate them. The spokesman of the Ministry of Foreign Affairs Oleg Nikolenko informed Suspilno about this. "The Ministry of Foreign Affairs maintains contact with Ukrainians who have expressed a desire to leave the Gaza Strip. Now there are 120 of them," Nikolenko informed. According to him, currently the embassies in Israel and Egypt, as well as the representative office of the Palestinian National Authority, are working out possible repatriation routes. The spokesman added that the Ministry of Foreign Affairs is keeping this situation under special control.	0.9000
120 громадян України хочуть вийхати з Сектору Гази через загострення на Близькому Сході. Посольства в Ізраїлі та Єгипті шукають способи їх евакуувати.		120 citizens of Ukraine want to leave the Gaza Strip because of the escalation in the Middle East. The embassies in Israel and Egypt are looking for ways to evacuate them.	

FIGURE 3.5: A summary-article pair from our custom dataset. The English version shows much higher score than the Ukrainian one.

texts in Ukrainian	BERTScore	texts in English	BERTScore
Через загострення ситуації на Близькому Сході з сектора Газа виявили бажання вийхати 120 громадян України. Наразі посольства в Ізраїлі та Єгипті шукають способи їх репатріації. Про це Суспільному повідомив речник МЗС Олег Ніколенко. "Міністерство закордонних справ підтримує зв'язок із українцями, які виявили бажання залишити Сектор Газа. Зараз їх 120 осіб", — поінформував Ніколенко. За його словами, наразі посольства в Ізраїлі та Єгипті, а також представництво при Палестинській національній адміністрації відпрацьовують можливі маршрути репатріації. Речник додав, що МЗС тримає цю ситуацію на особливому контролі.	0.5513	Due to the aggravation of the situation in the Middle East, 120 citizens of Ukraine expressed their desire to leave the Gaza Strip. Currently, the embassies in Israel and Egypt are looking for ways to repatriate them. The spokesman of the Ministry of Foreign Affairs Oleg Nikolenko informed Suspilno about this. "The Ministry of Foreign Affairs maintains contact with Ukrainians who have expressed a desire to leave the Gaza Strip. Now there are 120 of them," Nikolenko informed. According to him, currently the embassies in Israel and Egypt, as well as the representative office of the Palestinian National Authority, are working out possible repatriation routes. The spokesman added that the Ministry of Foreign Affairs is keeping this situation under special control.	0.7171
Горить нафтопровід на півдні Іраку. Нафтопостачання в тій частині країни припинене.		An oil pipeline is on fire in southern Iraq. Oil supply in that part of the country has been stopped.	

FIGURE 3.6: Summaries in summary-article pairs are swapped.

texts in Ukrainian	BERTScore	texts in English	BERTScore
За повідомленнями агенцій, у нафтовидобувному місті Кіркук від рук повстанців загинули один громадянин Південної Африки, один - Нової Зеландії та один іракець. За повідомленнями з Багдада, щонайменше шестеро людей, у тому числі четверо іноземців, зазнали поранень внаслідок вибуху. Вибух стався біля готелю в центрі міста, де переважно мешкають іноземні працівники. Вибух стався кількома годинами по тому, як принаймні семеро людей загинули і ще кілька десятків зазнали поранень внаслідок двох вибухів на одному зі столичних ринків. У кількох містах Іраку тривали сутички між силами коаліції та прихильниками шіїтського клерика Моктади аль-Садра. Американське командування повідомило про загибель понад 20 іракців, у тому числі кількох службовців поліції.	0.5685	According to news agencies, in the oil-producing city of Kirkuk, one citizen of South Africa, one citizen of New Zealand and one Iraqi died at the hands of the rebels. According to reports from Baghdad, at least six people, including four foreigners, were injured as a result of the explosion. The explosion occurred near a hotel in the city center, where mostly foreign workers live. The blast came hours after at least seven people were killed and dozens more injured in two explosions in one of the capital's markets. Clashes continued in several Iraqi cities between coalition forces and supporters of Shiite cleric Moqtada al-Sadr. The American command reported the death of more than 20 Iraqis, including several police officers.	0.8040
120 громадян України хочуть вийхати з Сектору Гази через загострення на Близькому Сході. Посольства в Ізраїлі та Єгипті шукають способи їх евакуувати.		120 citizens of Ukraine want to leave the Gaza Strip because of the escalation in the Middle East. The embassies in Israel and Egypt are looking for ways to evacuate them.	

FIGURE 3.7: Summaries in summary-article pairs are swapped.

Updated filtering criteria

After the first experiment was conducted and results analyzed, the decision was made to further filter the training data. Some of the summaries included long extracts from the text that fostered repetition of whole phrases and reduced abstractivity. In order to avoid this, a threshold on density score was imposed: density reflects squared length of extractive fragments and therefore highlights long consecutive excerpts. In addition, abstractivity threshold was applied and entries with low BERTScore were dropped.

The following filtering criteria were applied on top of the criteria mentioned in 3.3.1.

- density < 20
- abstractivity > 0.15
- BERTScore > 0.5

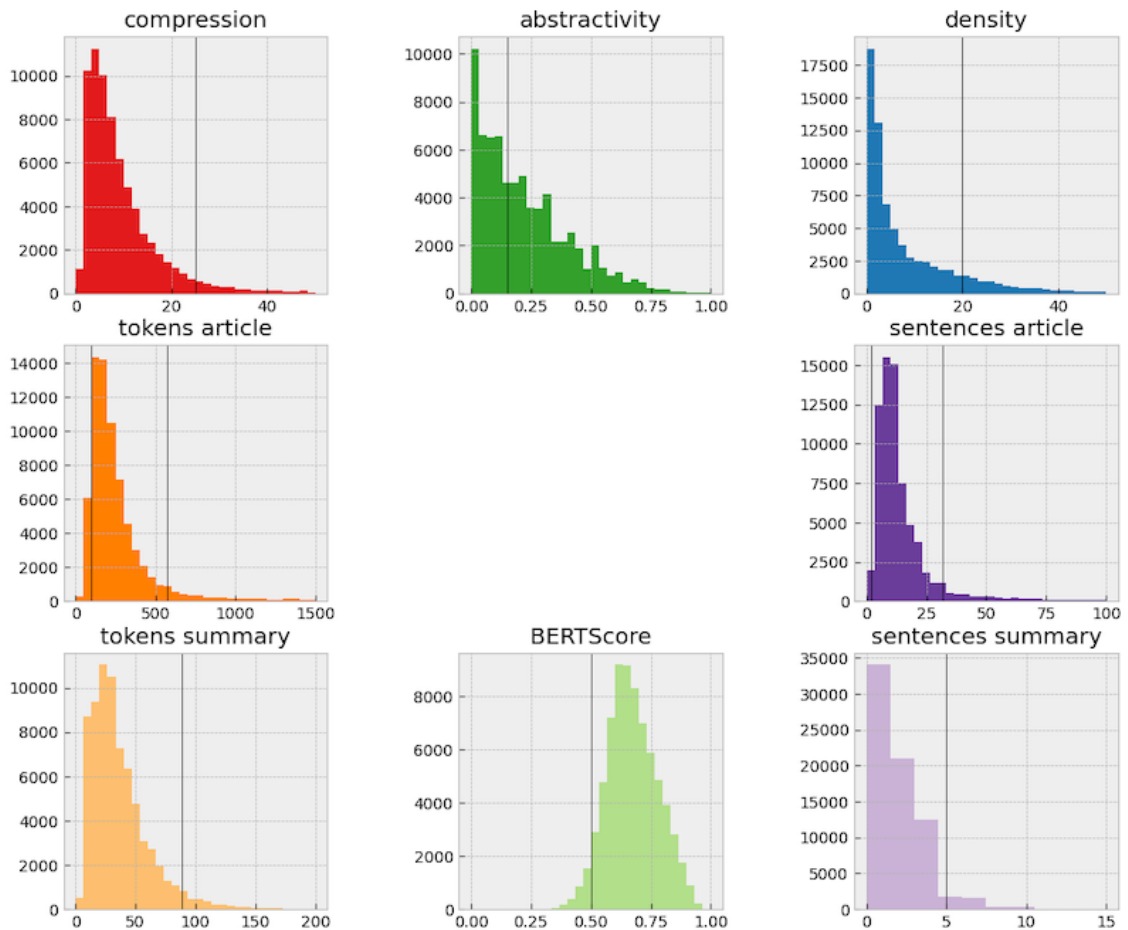


FIGURE 3.8: Distributions of these characteristics of the unfiltered dataset. Black lines represent the threshold of the second filtering (either lower or upper). These plots also do not include outliers.

Chapter 4

Approach

In this chapter we briefly describe main points of our experimental approach.

4.1 Fine-tuning

While LLMs are trained to be able to accomplish various natural language tasks, they often require some (or a lot) of additional fine-tuning. This is especially the case for less common languages like Ukrainian, which is supported only by a small number of available LLMs. However traditional fine-tuning is not practical or even feasible for models with number of parameters surpassing several billions. To address this problem, different methods of PEFT (Parameter-Efficient Fine-Tuning) [Mangrulkar et al., 2022] were suggested. The main idea lays in training a small number of extra parameters instead of re-training some of the existing ones, thus accomplishing the goal of improving model's performance for a specific task or data without utilizing enormous computational resources. Huggingface's documentation ¹ lists several options of PEFT methods available for causal language modelling: prompt-based tuning, LORA and IA3.

4.1.1 LORA

A popular way to efficiently train large models is to insert (typically in the attention blocks) smaller trainable matrices that are a low-rank decomposition of the delta weight matrix to be learnt during finetuning. The pretrained model's original weight matrix is frozen and only the smaller matrices are updated during training. This reduces the number of trainable parameters, reducing memory usage and training time which can be very expensive for large models.

LoRA (Low Rank Adaptation) [Hu et al., 2021] is a popular fine-tuning approach which involves using smaller trainable matrices that are a low-rank decomposition of the delta weight matrix while original weights of the model are freezed and unchanged. Subsequently those additional small matrices are injected into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. ²

Description of several of parameters in LORA configuration. Parameters descriptions in italics come from the documentation ³

`lora_r` - *Lora attention dimension (the "rank")*. Higher rank allows to train model more extensively but it also increases required processing time.

`lora_alpha` - *The alpha parameter for Lora scaling*.

¹<https://huggingface.co/docs/peft>

²https://huggingface.co/docs/peft/task_guides/lora_based_methods

³https://huggingface.co/docs/peft/v0.10.0/package_reference/lora#peft.LoraConfig

lora_dropout - The dropout probability for Lora layers.

target_modules - The names of the modules to apply the adapter to.

4.1.2 QLORA

As described by the authors of the paper [Dettmers et al., 2023], quantization is the process of discretizing an input from a representation that holds more information to a representation with less information. It often means taking a data type with more bits and converting it to fewer bits, for example from 32-bit floats to 8-bit Integers. QLORA (Quantized LORA) is a popular finetuning approach that drastically reduces memory usage thus enabling user to finetune large models while preserving task performance. QLORA backpropagates gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters (LoRA)[Dettmers et al., 2023].

4.2 Experimental setup

LLMs selection

The main focus of our experiments was on Llama-3 8B. It was released in April 2024. This model is the smaller one of the new update from the LLama family [Touvron et al., 2023]. According to its info page ⁴ it was trained on data in over 30 different languages (although predominantly English).

Text generation pipelines

Despite huggingface transformers having a "summarization" option that can be specified as a pipeline's task ⁵, it can not be used in our case: there is a limited list of models, that can be used for this option and neither of them is one of our target models. This necessitates using a "text-generation" task pipeline.

Hardware setup

To perform our experiments we utilized free-tier notebooks from kaggle ⁶, which allowed us to use GPU T4x2 for training. However, limitations included maximum training length of 12 hours, which prohibited some experiments from continuation.

⁴<https://ai.meta.com/blog/meta-llama-3/>

⁵https://huggingface.co/docs/transformers/v4.40.2/main_classes/pipelines

⁶<https://www.kaggle.com/>

Chapter 5

Experiments

In this section we describe experimental setup, evaluation metrics and analyze the results.

5.1 Zero-shot

Prompt selection

Selection of a good prompt can contribute to better results if such prompt was used during the initial training of the LLM. Otherwise the LLM can be fine-tuned to recognize and properly react to any prompt even if it was never used during the initial training. Such prompt selection for the summarization task in Ukrainian is not straightforward as there is no concise Ukrainian counterpart to phrase "summarize the following text:". There are possible alternatives, yet their meaning is not universal and using them may require additional explanation even to humans. Another approach that was tested includes combining an English prompt with text in Ukrainian, following the additional instruction to answer in Ukrainian. The results were below satisfactory and this approach was discarded. One of the examples illustrates zero-shot experiment for Llama-3 8B [5.2](#).

Therefore, the decision was made to select one of the prompts and train the model with it as none of the options proposed any meaningful pre-existing advantage. The prompt-answer format that was used for training and inference is presented in [5.1](#).

Analysis of results

From outputs of zero-shot experiments, it is clear that the model totally ignores any attempts to guide it into completing a specific task and just generates some text based on the input. Example [5.2](#) shows how the output consists of random generation or repetition.

```
f"""Напиши резюме тексту:  
{item["text"]}.  
Резюме: {item["summary"]}  
"""
```

FIGURE 5.1: The prompt-answer format that was used for training and inference.

<p>На чилійському острові Пасхи у Тихому океані пікап в'їхав в одну зі знаменитих старовинних статуй моаї і розтрощив її та платформу, на якій стояла скульптура. Про це заявив мер острова Педро Едмундса Паоа, пише газета El Mercurio. За його словами, аварія завдала надзвичайної шкоди статуї, а влада після інциденту хоче обмежити дорожній рух. «Всі були проти встановлення особливих правил дорожнього руху поблизу священних місць, але ми, місцева рада, пояснювали, яка небезпека існує» — розповів мер. Спільнота корінних мешканців острова опублікувала в Facebook допис з фотографіями з місця аварії і закликком захистити культурну спадщину острова. «Це не просто археологічний пам'ятник, це священний елемент живої культури, фундаментальний для світогляду рапануйців (корінного населення острова — ред.)», — йдеться в повідомленні. Кам'яні статуї моаї розташовані на узбережжі острова Пасхи. Вони виконані у формі людських фігур з величезними головами висотою до 20 метрів. Усього на острові Пасхи є 887 моаї, які датовані, ймовірно, 11 століттям.</p>	
"Напиши резюме тексту: "	"Summarize the following text: ", text, " Write answer in Ukrainian."
<p>Вважається, що їх створили корінні мешканці острова. Вважається, що статуї були створені для поклоніння богам, а також для захистення від ворогів. В 1995 році статуї моаї були внесені до списку Світової спадщини ЮНЕСКО.»\n\n## Посилання\n\n* Офіційний сайт [Архівовано 4 лютого 2011 у Wayback Machine.</p>	<p>Write answer in Ukrainian. Use the following words: аварія, авто, автошлях, автовідповідач, автовідповідач, автовідповідач, автовідповідач, автовідповідач, автовідповідач, автовідповідач, автовідповідач, автовідповідач, автовідповідач, автовідповідач, автовідповідач, автовідповідач</p>

FIGURE 5.2: Example of a zero-shot text summarization by Llama3 8B. The second row includes the prompts. It is clear that for both cases model generated possible continuation of the text with zero regards to attempts to ask it to do something specific.

5.2 Experiments

5.2.1 Llama finetuning

Base set of parameters

These parameters mainly remained the same along all the experiments. The ones that changed, will be noted later if applicable.

- lora_r = 128
- lora alpha = 32
- lora dropout = 0.05
- bias = "none"
- generation config.max new tokens = 120
- generation config.top p = 0.7
- generation config.repetition_penalty = 1.0
- optim="paged_adamw_8bit"
- lr_scheduler_type="cosine"
- warmup_ratio=0.01
- num trained epochs = 2

Experiment 1

The data used for training was a random sample of 10 000 from the pre-filtered larger dataset (first version of filtering criteria 3.3.1). Later for the purpose of obtaining quick results, another subset of 1000 out of 10 000 was selected.

training time : 2:05

Experiment 2

After examining the results of experiment 1, a second dataset filtering was performed 3.3.2. The parameters and dataset size remained the same for comparison.

training time : 2:02

Experiment 2.2

The same model that was trained in experiment 2, but during test inference in text generation parameters `generation_config.max_new_tokens = 220`.

Experiment 3

Training dataset was increased to 5000. Training stopped after 1 epoch due to training time exceeding maximum duration of 12 hours.

Experiment 4

Repetition penalty was set to 0.5, which was a mistake. Repetition got much worse.

Experiment 5

Repetition penalty was increased to 1.5, however problem with repetition did not get better regardless.

Prompt-answer format conformity

One of the unexpected behaviors during inference was that 2 epochs were apparently not enough for the model to adopt the expected output format as defined in 5.1 only about 50% of cases. The following is the count of correctly formatted outputs out of 100 generated summaries. In other cases, the output directly followed the input text and, therefore, looked as an additional paragraph that continued the article. Regardless of the format of the output, generated summaries were separated and analyzed. Probably this problem will disappear with more fine-tuning epochs and just indicates that the model still requires training.

- experiment 1: 62
- experiment 2: 47
- experiment 2.2: 50
- experiment 3: 58
- experiment 4: 21
- experiment 5: 26

5.2.2 Mistral finetuning

Unfortunately these experiments were unsuccessful. For some reason, Kaggle GPU option T4x2 was not working with Mistral models, which necessitates reserving to GPU P100. This resulted in much longer training time, which did not allow to fit even one epoch into the maximum length of training time on the website. The results did not improve even with lowered in half `lora_rank` parameter. Due to limited resources and time, this branch of experiments was closed.

5.2.3 Inference

For analysis of the generated results, a subset of 100 articles was selected from the summary-article pairs that were absent from training datasets. For the purpose of comparison, all models generated texts from same 100 articles. The small number of texts used for inference was determined by the need to analyze results piece-by-piece.

5.3 Evaluation

For evaluation of summary generation results, ROUGE-family metrics and BERTScore were selected. This choice was influenced both by feasibility of calculation for the Ukrainian language and by availability of existing results to compare with.

ROUGE-family metrics

ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. [Nenkova, 2011] For tokenization a library "tokenize-uk"¹ was used, for stemming "tree_stemmer" from ² was used.

$$ROUGE - N = \frac{\sum_{S \in \{Reference\ summaries\}} \sum_{gram_n} count_{match}(gram_n)}{\sum_{S \in \{Reference\ summaries\}} \sum_{gram_n} count(gram_n)} \quad (5.1)$$

With n standing for the length of the n-gram, $gram_n$, and $count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries.

$$F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2P_{lcs}} \quad (5.2)$$

$$R_{lcs} = \frac{LCS(X, Y)}{m} \quad (5.3)$$

$$P_{lcs} = \frac{LCS(X, Y)}{n} \quad (5.4)$$

BERTScore

BERTScore is an automatic evaluation metric for text generation. Analogously to common metrics, BERTSCORE computes a similarity score for each token in the candidate sentence with each token in the reference sentence. However, instead of exact matches, token similarity is computed using contextual embeddings. [Zhang et al., 2019]

For calculation of BERTScore for texts in Ukrainian, a version of Ukrainian Roberta "youscan/ukr-roberta-base"³ was used.

The results can be examined above. They appear to be quite low but the metrics appear to be a good relative indicator of which model performs worse. For example, the model in experiment 2.2 has a tendency to repeat clutter words and therefore it is understandable why it has the lowest scores. However, it is possible that n-gram based metrics are even less indicative for the Ukrainian language due to inflections of words.

In 5.1 scores show, that in combination these metric can be a good estimator of quality of generated summaries as higher scores (except with density) correlate with

¹<https://github.com/lang-uk/tokenize-uk>

²https://github.com/amakukha/stemmers_ukrainian/tree/master

³<https://huggingface.co/youscan/ukr-roberta-base>

На Сході окупанти намагаються утримувати позиції в районах населених пунктів Білогорівка, Бахмут, Веселого, Водяне, Времівка, Зайцеве, Золоте, Кліщівка, Кременна, Лиман, Нью-Йорк, Новодмитрівка, Новолуганське, Первомайське, Піски, Попасна, Роти, Соледар, Торецьк,

За попередньою інформацією, у районі залізничної станції в місті **Антрацит** Луганщини українські військові знищили значну частину озброєння та військової техніки РФ, яка транспортувалася залізничним транспортом. Кількість та характер ушкоджень уточнюються. Про це повідомив Генштаб ЗСУ. Упродовж поточної доби авіація Сил оборони завдала 13 ударів. Підтверджено ураження 12 районів зосередження озброєння та військової техніки, а також одного зенітного ракетного комплексу противника. Ракетні війська та артилерія уразили 3 пункти управління, 8 районів зосередження живої сили, озброєння і військової техніки та три склади боєприпасів. За уточненою інформацією, підтверджено знищення значної кількості особового складу противника за попередні дні. Так, у районах населених пунктів **Хлібодарівка** Донецької області, **Токмак** Запорізької області та **Токарівка** Херсонської області окупанти втратили до 150 осіб загиблими та близько 100 — пораненими. Також знищено 3 комплекси С-300 і до 10 одиниць іншого озброєння та військової техніки.

FIGURE 5.4: Generated summary(left) and input article (right). Settlement names are highlighted in the reference text. This case shows how the model can hallucinate but produces seemingly accurate statements as the locations mentioned in the summary all are valid places, located near the frontline in Donetsk oblast. The problem is only that they were not mentioned in the article.

За даними видання, Шольц не хоче, щоб Німеччина стала першою країною, яка надала Україні танки. Він вважає, що це може призвести до того, що інші країни також почнуть надавати танки Україні. Крім того, Шольц побоюється, що це може призвести до початку **третього світового конфлікту**. У той же час Шольц заявив, що Німеччина може надати Україні танки Leopard 2, якщо інші країни зроблять це раніше. **Однак інші країни, зокрема Велика Б**

Крім того, він заявив, що не бачить потреби в постачанні танків Україні, **оскільки вона вже отримала від Заходу достатньо сучасної зброї**. Крім того, Шольц заявив, що Німеччина не може постачати танки без згоди інших країн, які виробляють ці машини. Зокрема, це Франція, Велика Британія, Італія, Іспанія та Нідерланди. Однак, за словами Шольца, жодна з цих країн не погодилася

За словами джерел, Шольц заявив, що не бачить, чому Німеччина повинна надавати Україні танки, коли інші країни не роблять цього. Крім того, він заявив, що не бачить, чому Німеччина повинна надавати Україні танки, **коли інші країни не роблять цього**. За словами джерел, Шольц заявив, що не бачить, чому Німеччина повинна надавати Україні танки, коли інші країни не роблять цього. Крім того, він заявив, що не бачить

Значна частина громадян Німеччини проти постачання Україні своїх основних бойових танків Leopard 2, які просить Київ. Лише третина виступає за такі постачання. Як пише RND, про це свідчать результати опитування YouGov на замовлення DPA. Так, 45% опитаних висловилися проти надання Україні танків Leopard 2. Лише 33% висловилися за такі постачання, а ще 22% не мають конкретної думки щодо цього питання. Україна місяцями просить у Німеччини танки Leopard 2. Канцлер Олаф Шольц не хоче постачати їх самостійно, **вказуючи, що жодна інша країна не надала подібних систем зброї**. Водночас із боку його двох менших партнерів по коаліції, Зелених та ВДП, неодноразово лунали заклики все-таки відправити Leopard 2 до України. За це виступає й найбільша опозиційна фракція ХДС/ХСС. **Що передувало? Канцлер Німеччини Олаф Шольц раніше стверджував, що його країна надає достатньо допомоги Україні, а замість танків може зосередитися на постачанні артилерії та систем протиповітряної оборони**. Так він відповів на запитання, чому Берлін досі не надав Києву танків чи бойових літаків. Він пропонує розподіляти між партнерами завдання з допомоги Україні. Перед цим німецьке видання Welt із посиланням на джерела в українських урядових колах писало, що Олаф Шольц блокує постачання Україні 100 сучасних німецьких танків Leopard 2A7 навіть попри те, що компанія-виробник готова їх виготовити для українців. За даними видання, прем'єр-міністр України Денис Шмигаль під час візиту до Німеччини просив Шольца схвалити постачання Києву цих танків. Однак Шольц так і не взяв на себе зобов'язань стосовно схвалення постачання.

FIGURE 5.5: Three generated summaries from different models on the left and reference article on the right. Clear examples of hallucinations are highlighted. The messages delivered in the summaries are only partially related to the actual meaning of the article. However, even with little training, the models were able to produce results plausible enough to be confusing.

У селі Іванків Київської області окупанти спалили музей Марії Приймаченко, де зберігалися її роботи. Твори вдалося врятувати, повідомила художниця Таня Гончарова. "Російські окупанти спалили у смт Іванків історико-краєзнавчий музей, де зберігалися роботи відомої української народної художниці Марії Приймаченко. Твори Приймаченко іванківці вин

Російські війська знищили приміщення музею Марії Приймаченко, але твори вдалося врятувати — журналістка Таня Гончарова. Фото: Таня Гончарова. espresotb. espresotb. espresotb. espresotb. espresotb. espresotb. espresotb. espresotb. espresotb. espresotb. espresotb. espresotb.

Російські війська знищили приміщення музею Марії Приймаченко, але твори вдалося врятувати — журналістка Таня Гончарова. Фото: Таня Гончарова. espresotb.

Російські окупанти спалили у смт Іванків історико-краєзнавчий музей, де зберігалися роботи відомої української народної художниці Марії Приймаченко. Твори Приймаченко іванківці винесли по хатах. Вони не згоріли. Слава українцям. Музей відбудуємо. Фото: Таня Гончарова. espresotb.

Російські війська знищили приміщення музею, але твори вдалося врятувати. Російські війська знищили приміщення музею, але твори вдалося врятувати. Про це написала журналістка Таня Гончарова, інформує "Еспресо.Захід". "Російські окупанти спалили у смт Іванків історико-краєзнавчий музей, де зберігалися роботи відомої української народної художниці Марії Приймаченко. Твори Приймаченко іванківці винесли по хатах! Вони не згоріли!!!! Слава українцям!!!! Музей відбудуємо!!!!", - написала художниця. Нагадаємо, в неділю, 27 лютого, у смт Іванкові Київської області, окупанти знищили місцевий музей, де зберігалось близько 200 творів Марії Приймаченко. Усі були переконані, що картини були втрачені назавжди. Стежте за найважливішими новинами Львова, регіону, України та світу разом з "Еспресо.Захід"! Підписуйтесь на нашу facebook-сторінку.

FIGURE 5.6: These summaries (left) present several common problems: repetition and clutter. It also includes transformation of an emotional quote into a statement without corresponding attribution to the author.

clutter repetition. Changing the max new tokens parameter during fine-tuning also did not mitigate the issue.

- Repetition

In many cases, output contains repeated sentences or words. In some cases this repetitions are valid and meaningful, in others they are meaningless clutter. In either case, this is not a valid output. Possibly, this can be improved either with more training or with altering some lora or generation parameters.

- Clutter

The models tend to hallucinate non-existing photo descriptions, links and tags. This is easily explained by the training data because it was a deliberate choice not to clean this meta info related to telegram posts in order to train models to ignore this clutter. However either it was a bad idea or this type of behavior requires much more training or some change in parameters. In retrospect it is clear that more efforts should have been made to ensure removal of clutter, meta info and photo tags at least in the summaries. Its is possible that in such case the model would be less prone to hallucinations of this kind.

<p>Джоан Роулінг повернула собі лідерство у списку найбільш оплачуваних письменників світу за версією американського журналу Forbes. У 2019 році вона отримала дохід у \$92 мільйони. Джоан Роулінг повернула собі лідерство у списку найбільш оплачуваних письменників світу за версією американського журналу Forbes. У 2019 році вона отримала дохід у \$92 мільйони. Як пояснюють у Forbes, «фінансовим благословенням» для письменниці стала п'єса «Гаррі Поттер і прокляте дитя». У виданні зазначають, що заробити великі суми Роулінг вдалося завдяки продажу її книг та театральних квитків від постановки «Гаррі Поттера» у США. Друге місце у списку зайняв переможець минулого року, Джеймс Паттерсон, дохід якого склав \$70 млн. У 2018 році книга «Президент пропав», яку Паттерсон написав у співавторстві з Біллом Клінтоном, стала бестселером художньої літератури для дорослих. Колишня перша леді Мішель Обама замікає трійку лідерів з доходом приблизно у \$36 млн. Вона і її чоловік, колишній президент США Барак Обама, у 2017 році отримали, як повідомляється, аванс у розмірі \$65 млн за дві книги. «Книга Барака Обами ще не опублікована (ходять чутки, що вона може вийти під час виборів президента 2020 року), але успіх автобіографії Мішель Обами доводить, що для видавництва Penguin Random House захмарний аванс, можливо, окупився», — наголосили у Forbes. Нагадаємо, американський журнал Forbes опублікував рейтинг найбільш високооплачуваних письменників 2018 року. До нього увійшли 11 авторів. Так, найоплачуванішим письменником став американець Джеймс Паттерсон, який написав детективні романи «Невидимка», «Школа виживання», «І прийшов павук» і «Цілуєчи дівчат». Йому вдалося заробити 86 мільйонів доларів і просунути на другу сходинку лідерку 2017 року, авторку книг про пригоди Гаррі Поттера Джоан Роулінг з доходом 54 мільйони доларів. Трійку лідерів за доходом замикає «король жахів» Стивен Кінг, який заробив 21 мільйон доларів. Читайте також</p>	<p>Резюме: Джоан Роулінг повернула собі лідерство у списку найбільш оплачуваних письменників світу за версією Forbes. У 2019 році вона отримала дохід у 92 мільйони. Роулінг заробляла на продажах книг про Гаррі Поттера та театральних квитків від постановки «Гаррі Поттера» у США. Друге місце у списку зайняв Джеймс Паттерсон, дохід якого склав 70 мільйонів. Трійку лідерів замикає Мішель Обама, дохід якого склав 36 мільйонів. Вона написала автобіографію та отримала аванс за книгами, написані її колишнім чоловіком Бараком Обамом. Нагадаємо, у 2018 році Forbes опублікував рейтинг найбільш високооплачуваних письменників світу. Тодішній лідер рейтингу Джеймс Паттерсон заробив 86 мільйонів доларів. Друге місце посідала Джоан Роулінг, дохід письменниці склав 54 мільйони. Трійку лідерів замикає Стивен Кінг, який заробив 21 мільйон доларів.</p>
---	--

FIGURE 5.7: Example of a good summary that is simultaneously complete, well-formatted and factually correct.

<p>Один із найкращих результатів України в тенісі на Олімпіадах. Перша ракетка України Еліна Світоліна - у півфіналі одиночного турніру з тенісу, що проводиться на Олімпійських Іграх у Токіо! Українка у двох сетах впоралася з представницею Італії Камілою Джорджі. Українська тенісистка перед зустріччю з Джорджі провела три виснажливі поєдинки. Усіх трьох суперниць - Лауру Зігемунд, Айлу Томлянович та Марію Саккарі Світоліна змогла здолати лише в трьох сетах, причому в другому та третьому матчах українка здобувала вольові перемоги. У чвертьфіналі суперницею Світоліної стала представниця Італії Каміла Джорджі. Суперниці двічі зустрічалися між собою і здобули по одній перемозі, однак останній матч закінчився на користь італійки - у 2017 році на трав'яному турнірі в Бірмінгемі Джорджі здолала Світоліну. Проте українка не дозволила італійці повторити цей результат на Олімпіаді в Токіо. І завдала поразку Джорджі. У першому сеті Світоліна вийшла вперед 3:0, зробивши брейк на подачі Джорджі. Українці вдалося навіть збільшити перевагу до 5:1, але після цього Еліна програла три гейми поспіль. Проте партію не віддала - у десятому геймі українка використала четвертий сет-поінт та вийшла вперед. Активно розпочала Світоліна і другу партію. Ранній брейк на старті сету, після чого - ще один за рахунку 3:1. Проте у наступному ж геймі українка віддала гейм на власній подачі. Джорджі знову скоротила рахунок до 5:4, але до третього сету справа не дійшла. Світоліна взяла гейм на власній подачі і виграла другу партію. Попри перемогу, Світоліна припустилася 27 невимушених помилок, хоча й виконала шість ейсів. Каміла Джорджі (Італія) - Еліна Світоліна (Україна, 4) 0:2 (4:6, 4:6)</p>	<p>У півфіналі Світоліна зустрінеться з представницею США Софією Кенін, яка перемогла у чвертьфіналі Еліну Весніну з Росії. У фіналі зустрічатимуться переможниця матчу Світоліна - Кенін та переможниця матчу Джорджі - Кербер. Олімпійські ігри 2020. Теніс. Одиночний розряд. Чвертьфінал. Еліна Світоліна (Україна, 4) - Каміла Джорджі (Італія) 2:0 (6:4, 6:4).</p> <p>Світоліна вийшла на півфінал Олімпіади-2020 і зустрінеться там з представницею США Софією Кенін. У 2016 році Світоліна також вийшла на півфінал Олімпіади, але тоді програла Кенін. У 2020 році українка має шанс відігратися. Олімпіада-2020. Теніс. Одиночний турнір. Чвертьфінал. 1 серпня. Елі</p>
--	--

FIGURE 5.8: Example of two summaries that independently generate similar factually inconsistent information.

5.5 Interesting findings and conclusions

Example 5.8 shows that different models can generate similar information which is factually inconsistent with the reference text. For example, there is no mention of Sofia Kenin in the input text but her name is present twice in output texts. The most interesting thing is that her name is not present in the subset on which the models trained either, however the match between Svitolina and Kenin happened in reality in 2023 (while the text in this example is supposed to be about 2020).

Example 5.7 shows how models can provide correct summaries by paraphrasing or removing some sentences. It is interesting that throughout this text all numbers are valid and consistent with the reference text.

По обіді 15 червня війська Російської Федерації з системи залпового вогню "Ураган" обстріляли місто Зеленодольськ Криворізького району Дніпропетровщини. Про це повідомив **начальник Криворізької районної військової адміністрації Євген Ситниченко**. За його даними, у результаті обстрілу пошкоджено кілька будинків. Загиблих та поранених серед мирного населення немає, сказав Ситниченко. Пізніше у своєму телеграмі **голова Дніпропетровської обласної військової адміністрації Валентин Резніченко** уточнив, що війська РФ з «Ураганів» вдарили по Зеленодольську та Апостоловому. У Зеленодольську поцілили по житлових кварталах: пошкоджені кілька багатоповерхівок. Попередньо, люди не постраждали. "Місця обстрілів біля будинків засіяні касетними боеприпасами. Працюють піротехніки", – написав він. В Апостоловому – займання в лісосмузі. Рятувальники гасять пожежу.

Резюме: Внаслідок обстрілу міста Зеленодольськ на Дніпропетровщині пошкоджено кілька будинків, повідомив **начальник Криворізької РВА Євген Ситниченко**. За його словами, загиблих та поранених серед мирного населення немає. Пізніше **голова Дніпропетровської ОВА Валентин Резніченко** уточнив, що війська РФ з «Ураганів» вдарили по Зеленодольську та Апостоловому.

FIGURE 5.9: Example showcasing ability to make correct abbreviations without losing context.

Another interesting aspect 5.9 highlights how models are able to identify specific terms and make suitable abbreviations without missing important parts of information. This is an example concerning military administration, however there are many other specifically relevant to Ukraine terms like abbreviations of church names etc. Maybe this a consequence of data on which Llama-3 was initially trained or this behavior was picked up on during fine tuning due to proliferation of such abbreviation patterns in the training dataset.

Chapter 6

Conclusions and discussion

In this master thesis we conducted a review of relevant literature and existing datasets, outlined problems with existing data, developed a plan for creation of a custom dataset, gathered data, implemented a fine-tuning and generation pipeline, conducted several experiments and analyzed the results both quantitatively and qualitatively.

The preliminary results show that reasonably satisfactory results can be achieved even with little fine tuning (LLaMA3 8b + 2 epochs with QLoRA on small training subset and even 1 epoch with larger training subset). However, they also show several important problems, that require solution or at least mitigation. For example, there is a problem with repetitions and incomplete sentences. In addition, there is a problem that is typical for all content generated with LLMs – hallucinations or factual inconsistencies. Some problems may be mitigated by further cleaning the training dataset, which contains some clutter info that is reproduced in the generated texts.

Further research is needed to find better combination of parameters and also to fine-tune with more computational resources, which would allow more epochs, larger datasets and hopefully better subsequent performance of a model.

All relevant code and datasets can be accessed on GitHub ¹.

¹<https://github.com/JuliaWozniak/Abstractive-summarization-Ukrainian>

Bibliography

- Aharoni, Roei et al. (2023). “Multilingual Summarization with Factual Consistency Evaluation”. en. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 3562–3591. DOI: [10.18653/v1/2023.findings-acl.220](https://doi.org/10.18653/v1/2023.findings-acl.220). URL: <https://aclanthology.org/2023.findings-acl.220> (visited on 03/22/2024).
- Ahvir, Vicent et al. (2024). “Abstractive Summarizers Become Emotional on News Summarization”. In: *Applied Sciences* 14.2. ISSN: 2076-3417. DOI: [10.3390/app14020713](https://doi.org/10.3390/app14020713). URL: <https://www.mdpi.com/2076-3417/14/2/713>.
- Akani, Eunice et al. (2023). “Reducing named entity hallucination risk to ensure faithful summary generation”. en. In: *Proceedings of the 16th International Natural Language Generation Conference*. Prague, Czechia: Association for Computational Linguistics, pp. 437–442. DOI: [10.18653/v1/2023.inlg-main.33](https://doi.org/10.18653/v1/2023.inlg-main.33). URL: <https://aclanthology.org/2023.inlg-main.33> (visited on 03/18/2024).
- Bommasani, Rishi and Claire Cardie (2020). “Intrinsic Evaluation of Summarization Datasets”. en. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8075–8096. DOI: [10.18653/v1/2020.emnlp-main.649](https://doi.org/10.18653/v1/2020.emnlp-main.649). URL: <https://www.aclweb.org/anthology/2020.emnlp-main.649> (visited on 03/31/2024).
- Cohan, Arman et al. (2018). *A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents*. arXiv: [1804.05685](https://arxiv.org/abs/1804.05685) [cs.CL].
- Dettmers, Tim et al. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv: [2305.14314](https://arxiv.org/abs/2305.14314) [cs.LG].
- Devlin, Jacob et al. (2018). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: Publisher: [object Object] Version Number: 2. DOI: [10.48550/ARXIV.1810.04805](https://doi.org/10.48550/ARXIV.1810.04805). URL: <https://arxiv.org/abs/1810.04805> (visited on 05/09/2024).
- Fabbri, Alexander R. et al. (2019). *Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model*. arXiv: [1906.01749](https://arxiv.org/abs/1906.01749) [cs.CL].
- Fabbri, Alexander R. et al. (2020). “SummEval: Re-evaluating Summarization Evaluation”. In: Publisher: [object Object] Version Number: 4. DOI: [10.48550/ARXIV.2007.12626](https://doi.org/10.48550/ARXIV.2007.12626). URL: <https://arxiv.org/abs/2007.12626> (visited on 05/09/2024).
- Galeshchuk, Svitlana (2023). “Abstractive Summarization for the Ukrainian Language: Multi-Task Learning with Hromadske.ua News Dataset”. en. In: *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 49–53. DOI: [10.18653/v1/2023.unlp-1.6](https://doi.org/10.18653/v1/2023.unlp-1.6). URL: <https://aclanthology.org/2023.unlp-1.6> (visited on 03/18/2024).
- Goyal, Tanya, Junyi Jessy Li, and Greg Durrett (2023). *News Summarization and Evaluation in the Era of GPT-3*. arXiv: [2209.12356](https://arxiv.org/abs/2209.12356) [cs.CL].
- Grusky, Max, Mor Naaman, and Yoav Artzi (2018). “Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies”. In: Publisher: [object Object] Version Number: 2. DOI: [10.48550/ARXIV.1804.11283](https://doi.org/10.48550/ARXIV.1804.11283). URL: <https://arxiv.org/abs/1804.11283> (visited on 04/05/2024).

- Hasan, Tahmid et al. (2021). “XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages”. In: Publisher: [object Object] Version Number: 1. DOI: [10.48550/ARXIV.2106.13822](https://doi.org/10.48550/ARXIV.2106.13822). URL: <https://arxiv.org/abs/2106.13822> (visited on 03/26/2024).
- Hu, Edward J. et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv: [2106.09685](https://arxiv.org/abs/2106.09685) [cs.CL].
- Kryściński, Wojciech et al. (2019). “Evaluating the Factual Consistency of Abstractive Text Summarization”. In: Publisher: [object Object] Version Number: 1. DOI: [10.48550/ARXIV.1910.12840](https://doi.org/10.48550/ARXIV.1910.12840). URL: <https://arxiv.org/abs/1910.12840> (visited on 04/05/2024).
- Ladhak, Faisal et al. (2020). *WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization*. arXiv: [2010.03093](https://arxiv.org/abs/2010.03093) [cs.CL].
- Lavie, Alon and Abhaya Agarwal (July 2007). “METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments”. In: pp. 228–231.
- Lin, Chin-Yew and Eduard Hovy (2003). “Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics”. In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 150–157. URL: <https://aclanthology.org/N03-1020>.
- Liu, Yang and Mirella Lapata (2019). “Text Summarization with Pretrained Encoders”. In: Publisher: [object Object] Version Number: 2. DOI: [10.48550/ARXIV.1908.08345](https://doi.org/10.48550/ARXIV.1908.08345). URL: <https://arxiv.org/abs/1908.08345> (visited on 04/01/2024).
- Luo, Zheheng, Qianqian Xie, and Sophia Ananiadou (2023). “ChatGPT as a Factual Inconsistency Evaluator for Text Summarization”. In: Publisher: [object Object] Version Number: 2. DOI: [10.48550/ARXIV.2303.15621](https://doi.org/10.48550/ARXIV.2303.15621). URL: <https://arxiv.org/abs/2303.15621> (visited on 03/18/2024).
- Mangrulkar, Sourab et al. (2022). *PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods*. <https://github.com/huggingface/peft>.
- Maynez, Joshua et al. (2020). “On Faithfulness and Factuality in Abstractive Summarization”. en. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1906–1919. DOI: [10.18653/v1/2020.acl-main.173](https://doi.org/10.18653/v1/2020.acl-main.173). URL: <https://www.aclweb.org/anthology/2020.acl-main.173> (visited on 03/18/2024).
- Nallapati, Ramesh et al. (2016). “Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond”. en. In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, pp. 280–290. DOI: [10.18653/v1/K16-1028](https://doi.org/10.18653/v1/K16-1028). URL: <http://aclweb.org/anthology/K16-1028> (visited on 03/31/2024).
- Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (2018a). “Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, pp. 1797–1807. DOI: [10.18653/v1/D18-1206](https://doi.org/10.18653/v1/D18-1206). URL: <https://aclanthology.org/D18-1206>.
- (2018b). “Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, pp. 1797–1807. DOI: [10.18653/v1/D18-1206](https://doi.org/10.18653/v1/D18-1206). URL: <https://aclanthology.org/D18-1206>.
- Nenkova, Ani (2011). “Automatic Summarization”. en. In: *Foundations and Trends® in Information Retrieval* 5.2, pp. 103–233. ISSN: 1554-0669, 1554-0677. DOI: [10.1561/](https://doi.org/10.1561/)

1500000015. URL: <http://www.nowpublishers.com/article/Details/INR-015> (visited on 03/31/2024).
- Papineni, Kishore et al. (2001). “BLEU: a method for automatic evaluation of machine translation”. en. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Philadelphia, Pennsylvania: Association for Computational Linguistics, p. 311. DOI: [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135). URL: <http://portal.acm.org/citation.cfm?doid=1073083.1073135> (visited on 03/30/2024).
- Paulus, Romain, Caiming Xiong, and Richard Socher (2017). “A Deep Reinforced Model for Abstractive Summarization”. In: Publisher: [object Object] Version Number: 3. DOI: [10.48550/ARXIV.1705.04304](https://doi.org/10.48550/ARXIV.1705.04304). URL: <https://arxiv.org/abs/1705.04304> (visited on 04/01/2024).
- Raffel, Colin et al. (2023). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. arXiv: [1910.10683](https://arxiv.org/abs/1910.10683) [cs.LG].
- Rush, Alexander M., Sumit Chopra, and Jason Weston (2015). “A Neural Attention Model for Abstractive Sentence Summarization”. In: Publisher: [object Object] Version Number: 2. DOI: [10.48550/ARXIV.1509.00685](https://doi.org/10.48550/ARXIV.1509.00685). URL: <https://arxiv.org/abs/1509.00685> (visited on 04/01/2024).
- Scialom, Thomas et al. (2020). “MLSUM: The Multilingual Summarization Corpus”. en. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8051–8067. DOI: [10.18653/v1/2020.emnlp-main.647](https://doi.org/10.18653/v1/2020.emnlp-main.647). URL: <https://www.aclweb.org/anthology/2020.emnlp-main.647> (visited on 04/01/2024).
- See, Abigail, Peter J. Liu, and Christopher D. Manning (2017). “Get To The Point: Summarization with Pointer-Generator Networks”. In: Publisher: [object Object] Version Number: 2. DOI: [10.48550/ARXIV.1704.04368](https://doi.org/10.48550/ARXIV.1704.04368). URL: <https://arxiv.org/abs/1704.04368> (visited on 04/01/2024).
- Touvron, Hugo et al. (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: [2302.13971](https://arxiv.org/abs/2302.13971) [cs.CL].
- Varab, Daniel and Natalie Schluter (2021). “MassiveSumm: a very large-scale, very multilingual, news summarisation dataset”. en. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10150–10161. DOI: [10.18653/v1/2021.emnlp-main.797](https://doi.org/10.18653/v1/2021.emnlp-main.797). URL: <https://aclanthology.org/2021.emnlp-main.797> (visited on 03/18/2024).
- Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems*, pp. 5998–6008. URL: <http://arxiv.org/abs/1706.03762>.
- Zhang, Haopeng, Xiao Liu, and Jiawei Zhang (2023). *SummIt: Iterative Text Summarization via ChatGPT*. arXiv: [2305.14835](https://arxiv.org/abs/2305.14835) [cs.CL].
- Zhang, Jingqing et al. (2020). “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 11328–11339. URL: <https://proceedings.mlr.press/v119/zhang20ae.html>.
- Zhang, Tianyi et al. (2019). “BERTScore: Evaluating Text Generation with BERT”. In: Publisher: [object Object] Version Number: 3. DOI: [10.48550/ARXIV.1904.09675](https://doi.org/10.48550/ARXIV.1904.09675). URL: <https://arxiv.org/abs/1904.09675> (visited on 04/11/2024).