PersonalSum: A User-Subjective Guided Personalized Summarization Dataset for Large Language Models

Lemei Zhang Peng Liu* Marcus Tiedemann Oekland Henriksboe
Even W. Lauvrak Jon Atle Gulla Heri Ramampiaro
Department of Computer Science, Norwegian University of Science and Technology
{lemei.zhang, peng.liu, jon.atle.gulla, heri}@ntnu.no

Abstract

With the rapid advancement of Natural Language Processing in recent years, numerous studies have shown that generic summaries generated by Large Language Models (LLMs) can sometimes surpass those annotated by experts, such as journalists, according to human evaluations. However, there is limited research on whether these generic summaries meet the individual needs of ordinary people. The biggest obstacle is the lack of human-annotated datasets from the general public. Existing work on personalized summarization often relies on pseudo datasets created from generic summarization datasets or controllable tasks that focus on specific named entities or other aspects, such as the length and specificity of generated summaries, collected from hypothetical tasks without the annotators' initiative. To bridge this gap, we propose a high-quality, personalized, manually annotated abstractive summarization dataset called PersonalSum. This dataset is the first to investigate whether the focus of public readers differs from the generic summaries generated by LLMs. It includes user profiles, personalized summaries accompanied by source sentences from given articles, and machine-generated generic summaries along with their sources. We investigate several personal signals — entities/topics, plot, and structure of articles—that may affect the generation of personalized summaries using LLMs in a few-shot in-context learning scenario. Our preliminary results and analysis indicate that entities/topics are merely one of the key factors that impact the diverse preferences of users, and personalized summarization remains a significant challenge for existing LLMs. Our dataset and code are available at https://github.com/SmartmediaAI/PersonalSum.

1 Introduction

Recent studies have demonstrated significant improvements in generating generic summaries using Large Language Models (LLMs) like GPT-3.5, achieving state-of-the-art, even human-level, performance on standard summarization benchmarks. However, personalized summarization, which condenses text to match user preferences while maintaining relevance and non-redundancy remains largely unexplored.

Kukoleva *et al.* [1] identified three distinct user reading habits in the news domain: **Attentive reading**, where users read the full article attentively, focusing on details, **Selective reading**, where users focus only on interesting fragments; and **Scanning**, where users absorb only the important ideas. User reading focus and attentive reading time can be achieved by monitoring reading duration and scrolling depth, or asking users to annotate their interested parts explicitly. However, there are no existing publicly available resources for research purposes in this area.

*Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/079017-3151

Table 1: Comparison between PersonalSum and existing popular summarization datasets.

Datasets	Language	Domain	#Summaries	Construction		User Profile	Summary Source	Personalized
				Human Annotation	Multi-annotation			
Generic Summarization Datasets								
CNN/DM [2]	English	News	311,971	√	×	×	×	×
XSum [3]	English	News	226,711	√	×	×	×	×
NewsRoom [4]	English	News	1,212,740	√	×	×	×	×
BigPatent [5]	English	Academic	1,341,362	×	×	×	×	×
arXiv [6]	English	Academic	215,913	×	×	×	×	×
PubMed [6]	English	Academic	133,215	×	×	×	×	×
LCSTS [7]	Chinese	News	2,400,591	√	×	×	×	×
WikiHow [8]	English	WikiHow	230,843	×	×	×	×	×
Controllable Summarization Datasets								
DUC [9]	English	News	300	√	√	×	×	×
QMSum [10]	English	Meetings	1,808	√	√	×	×	×
WikiAsp [11]	English	Wikipedia	566,881	×	√	×	×	×
MACSUM [12]	English	News&Meetings	8333	√	√	×	√	×
Personalized Summarization Datasets								
Amazon Reviews [13]	English	E-commerce	571,540,000	√	√	√	×	√
PENS [14]	English	News Headline	20,600	√	√	×	×	√
PersonalSum (ours)	Norwegian	News	1,816	√	√	√	√	√

Existing textual summarization datasets still suffer from several limitations. First, most data annotators are predominantly journalists or professional writers in related fields, or they are few in number (e.g. only single-digit annotators), resulting in a lack of representativeness, personalization and diversity of the annotated summaries for the general public. Apart from that, the lack of crucial user information, such as reading time and specific content engagement, limits existing research to generic and controllable summary generation. Including user-specific data could enable more personalized and user-centric studies. Second, existing summarization datasets do not include the annotators' user profiles, making our work the only available data for personalized summarization tasks at this stage.

To this end, we propose PersonalSum, a high-quality human-annotated dataset for personalized news summarization with multiple attributes. In PersonalSum, each article includes personalized summaries annotated by multiple ordinary users based on their interests, multiple pairs of questions and answers related to news articles, and generic summaries generated by machines with manual proofreading. Both the summaries and the question-and-answer pairs contain source information corresponding to the original text. We validated the personalized nature of the dataset and compared it with machine-generated summaries. Today, where LLMs are widely used for text summarization, collecting personalized summaries driven by user subjectivity is challenging. Moreover, we cannot verify the effectiveness of manual annotations by comparing their similarity to machine-generated summaries. Our analysis revealed that most machine-generated summaries tend to rephrase the introductory sections of news articles, which may also align with user interests. To ensure good annotation quality, we designed an iterative approach combining human evaluation and LLM outputs. Based on user subjectivity, we preliminarily explored the capability of LLMs to generate personalized summaries under different in-context learning settings. Inspired by the findings by Kukoleva *et al.* [1] for determining user interest points, we investigated the impact of entities/topics, plot, and article structure on extracting personalized summaries. We have made the above-mentioned datasets, codes and documents available at `https://github.com/SmartmediaAI/PersonalSum` under a CC BY-NC 4.0 license.

2 Related work

The personalized textual summarization task is an important yet challenging area in Natural Language Processing. It aims to generate concise and targeted summaries based on a user's personal preferences and focus. However, due to the lack of publicly available datasets, there has been very little research that has investigated this problem. Currently, publicly available summarization datasets can roughly be divided into three different categories. The first category involves the generation of generic summaries. These summaries can be either pseudo datasets, created by extracting the abstract of the article [15, 18, 19] or the highlighted sentences within each paragraph [16, 21], or they can be datasets annotated by professional writers or journalists, leveraging their expertise [17, 20].

Although generic text summarization methods are useful for general-purpose summarization tasks, they frequently fall short in meeting the specific intentions and needs of individual users. This shortcoming has prompted the development of the second category of datasets, called the Controllable Text Summarization (CTS) techniques, an expanding corpus of research dedicated to this area [25]. The task of CTS focuses on creating summaries of source documents that adhere to various controllable attributes or aspects, such as summary length and coverage of specific topics [11, 12, 26, 27, 28]. The primary distinction between controllable summarization datasets and PersonalSum lies in the method of controlling the summaries. The former employs explicitly given attributes to guide the generation of summaries, ensuring that these control factors are clearly present in the dataset. In contrast, the personalized news summary dataset proposed in this paper is based on users annotating points of interest within the articles they read. The personalized factors in this dataset are implicit, making them more closely aligned with user interests as reflected in real-world scenarios.

The third category of datasets for textual summarization (e.g. Amazon Reviews [13]) focuses on generating personalized reviews by considering a series of discrete attributes of a given product, which is often an important part of recommender systems [15, 16, 17, 18, 19, 20, 21, 22, 23, 24]. Different from generic summarization datasets, this dataset category usually contains various personalized information, e.g., ratings, user and product IDs, and history text, etc. The difference between personalized review summarization and personalized news summarization is that the former often contains user sentiment information. Further, in specific domains, the vocabulary in user reviews tends to be relatively limited. For example, the vocabulary used to describe a movie or a song is not typically used to describe an electronic product, and vice versa. News summaries, on the other hand, often focus on facts and are expressed in a relatively neutral manner, rarely incorporating personal emotions.

To the best of our knowledge, the most similar dataset to this paper is called PENS, which is used for personalized headline generation [14]. Headline generation is a special case of abstractive text summarization where the goal is to create a concise, often single-sentence "summary" highlighting one key fact of a news article. In contrast, the summaries in PersonalSum may contain two or more user interested points that reflect different perspectives on the news. Furthermore, in PersonalSum, the sources, a feature that is absent in the other two personalized summarization datasets, can be utilized for various analyses, such as reflecting the user's information extraction habits, serving as a reference for model interpretability, and extracting user interests. A comparison among representative summarization datasets is shown in Table 1.

3 PersonalSum

In this section, we detail our PersonalSum dataset from the perspective of data preparation and collection, while ensuring annotation quality for personalized summarization.

3.1 Dataset construction

The data collection process requires iterative efforts, primarily involving 3 stages, as detailed below.

Stage 1: Construction of generic summaries. In the initial stage, we carefully selected 465 news articles evenly distributed across 10 categories, provided by Schibsted[2], an international media company. We then developed a data annotation platform and recruited three Norwegian native-speaking students for this task. The students were instructed to revise the GPT-4 generated summaries, emphasizing language fluency, factual consistency, and coherence with the given article while maintaining the focus of the machine-generated summaries. Meanwhile, the students were also asked to highlight sentences from the given news article indicating sources of the summary. After that, each summary was cross-checked by two distinct students with 100% internal agreement. In addition to this, each article was paired with multiple question-answer (QA) sets related to the article content, serving as a component of the quality control process during the second phase.

Stage 2: Construction of personalized summaries. In this stage, we recruited annotators from Amazon Mechanical Turk for personalized summarization tasks. First, we instituted a qualification process to assess the suitability of workers for the annotation task. A questionnaire, encompassing fluency in Norwegian, demographic information, news consumption habits, areas of interest, and

[2]https://schibsted.com/

https://doi.org/10.52202/079017-3151

gender, was administered, allowing us to filter potential workers based on different criteria. Workers without fluent Norwegian knowledge were filtered out. To ensure that each user has certain historical data for later analysis and to ensure the diversity of annotations per article, we then decided to include three articles in each HIT[3], with each HIT assigned to at least three different annotators. Detailed instructions were provided to guide workers through the annotation process, emphasizing the creation of concise and informative summaries aligned with the annotator's own preferences and interests, while also providing the source of the summary in the given article. Further, workers were instructed to choose the correct question-answer pair related to the given article from three options to assess comprehension and validate annotations. HITs without a passing rate of 2/3 on the single-choice questions were automatically rejected. Other automatic filtering rules include that the source should come directly from the article, the summary length should not be less than 50 words, and the task duration should not be less than 5 minutes based on the statistics of the time peoples used per reading session in [1], among others, to pre-control the data quality. Considering the market salary level, annotators were compensated $6 USD per approved HIT, reflecting an hourly rate of $18 USD, with a projected completion time of 30 minutes per task.

Stage 3: Post quality control. After checking randomly sampled data, we found that a considerable proportion of the collected summaries did not meet the collection standards. They either lack semantic coherence, are irrelevant to the given source, or are annotated in other languages such as English. Inspired by recent work that leverages LLMs as evaluators [27], we utilized OpenAI GPT-3.5-Turbo with few-shot prompting to evaluate the annotated summaries from the perspectives of *Coherence*, *Consistency*, and *Relevance*. In addition to evaluating the coherence of summaries and news articles based on existing work, we especially evaluated the relevance between summaries and sources. Specifically, we score relevance from 0 to 1 according to the relevant scale between the summary and its source. Annotations with a relevance score of less than 0.8 were set aside for human evaluation. For the remaining annotations with a relevance score greater than 0.8, a random sample of 10 percent was taken for human evaluation to estimate the accuracy of the LLM results. This is a very effective way to control the annotation quality. One concern raised during the post-quality control process is that certain sources include only a part of the content of the summary, rather than the entirety. This situation is commonly seen when the relevant score is greater than 0.8. In this context, we reached a consensus that as long as the source effectively conveys the main points within the summary, it is considered qualified. After the second phase of the first round, only 668 annotations out of 1,395 met the collection criteria. Based on a spot check of the evaluation results, we found that GPT-3.5-Turbo can achieve 98% accuracy for data with a relevance score greater than 0.8, and 95% accuracy for evaluations less than 0.8. We then iterated through stages 2 and 3 until the predefined evaluation is satisfied or we reach the budget limit. The prompt used for evaluation is shown in the supplementary material. We use one-shot prompting to control the format of the returned results.

3.2 Statistical analysis

This section introduces the basic statistics of PersonalSum, covering: 1) annotators, articles, and summaries; 2) the differences between personalized and machine-generated summaries.

After the data collection process described in Section 3.1, we collected a total of 1099 personalized summaries from 441 news articles annotated by 39 distinct Amazon Turkers. Figure 1(a)-1(d) shows the distributions of the annotators based on their gender, age, reading habits and occupation. These distributions indicate that the annotators come from diverse demographic groups, allowing the annotated summaries to represent the perspectives of ordinary people to some extent. Figure 1 (e)-1(g) present basic statistics from the article perspectives. In Figure 1(h) and 1(i), we present the distributions of qualified annotated summaries per worker and the time consumed per annotation.

Figure 2(b) shows that machine-generated summaries are relatively longer than human-annotated ones, but we believe this should not be a major difference since LLMs can generate shorter summaries with different sampling strategies. To illustrate the distribution of summary sources within its given news article, we evenly divided the article into five parts based on the number of sentences, labeling them 1, 2, 3, 4, and 5 from the beginning to the end. Each sentence's source was then identified according to its corresponding part of the article. If multiple sentences originate from the same part, they were counted only once. In Figure 2(a), it can be seen that a considerable number of machine-

[3]In our task, a HIT is one task assigned to an annotator that contains three qualification tests and three different news articles for annotation.

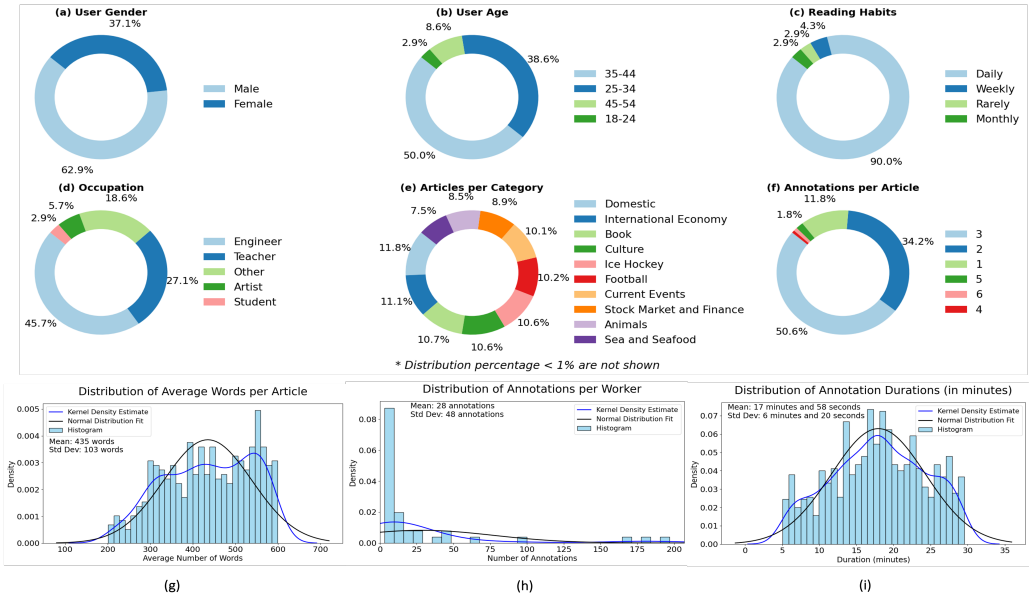

Figure 1: (a)-(d) shows annotator demographics, including gender, age, reading habits, and occupation. (e)-(g) cover annotation categories and counts. (h) and (i) display the distribution of qualified summaries per worker and time spent per annotation.

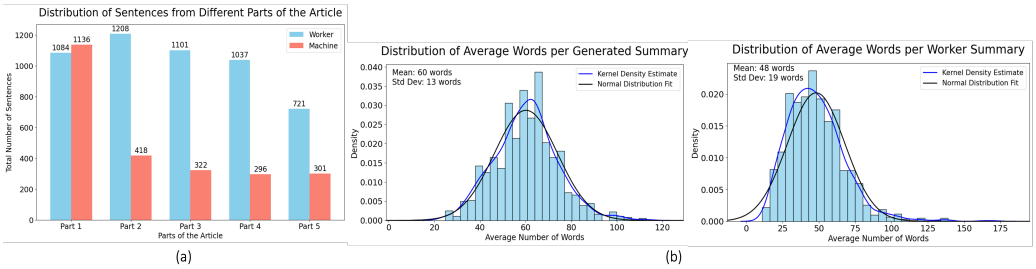

Figure 2: (a) The distribution of sources of machine-generated summaries and human-annotated personalized summaries. (b) The distribution of average words per machine-generated summary and human-annotated summary.

generated summaries originate from the first part of the article. In contrast, the manually annotated summaries are relatively evenly distributed across various parts of the article. This discrepancy may arise because most existing summary datasets use the abstract, typically located at the beginning of the news article, as the ground truth. This observation also indicates that users' focus is diverse and not limited to the initial section of the article. Further, we analyzed the distribution of different annotation sources for the same article. Out of 441 articles, 52 had only one annotation, 36 had different user summaries originating from the same article part(s), and 353 had annotations from different parts of the article. Notably, the sources for these 36 articles were inconsistent at the sentence level.

4 Experiments

4.1 Models and evaluation metrics

Considering the limitations of language and data size in PersonalSum, we tested our dataset on four LLMs across different architectures and model scales, namely OpenAI GPT-3.5 Turbo[4], Llama3-

[4]https://platform.openai.com/docs/models/gpt-3-5-turbo

https://doi.org/10.52202/079017-3151

instruct[5], Google Gemini-1.0-pro[6], and NorwAI-Mixtral-8x7B-instruct[7]. All of the selected models are reported to support Norwegian prompting and complex tasks. We benchmarked GPT-3.5-Turbo, Gemini-1.0-pro and NorwAI-Mixtral-8x7B-instruct in zero-shot (for generating generic summaries) and 2/5/10-shot prompting with Norwegian prompts[8]. For all tested LLMs in this paper, we sampled with a temperature of 0.3, following the work of Wu *et al.*[29]. We evaluated the Llama3-Instruct model using 2-shot prompting and observed that, aside from producing generic summaries, it often generated English summaries for longer Norwegian prompts. This led to significantly lower test results compared to generic summaries. The subpar performance may be due to the limited Norwegian data in the Llama3 pre-training dataset or suboptimal prompt design. Therefore, we chose not to test it with other prompt settings.

We present the performance of PersonalSum on several representative summarization metrics: ROUGE-1/2/L [30], and BERTScore (F1) [31][9]. Inspired by Maynez et al.[32], who used an entailment score to measure the factuality of generated text compared to human-written text in abstractive summarization, we adopted an entailment model [33][10] pretrained on NorBERT [34] to assess the entailment relationship between the generated summaries and the human-annotated personalized summaries in the PersonalSum dataset. Specifically, we consider a generated summary to successfully capture the annotator's interests if it entails any part of the human-written summary for the corresponding news article. Based on the same work, we calculated the entailment score as the summation of the ratio of support and neutral, as the contradict ratio means that the generated summaries violate the human-written golden summaries. Due to page limits, we only report the best performance of different promptings. The final performances are reported as the average of test results over 5 runs. For the complete evaluation results, please refer to the supplementary materials.

4.2 Data preprocessing and prompting

To investigate the impact of three factors, entity/topics, story plots, and article structure, on users' personalized summarization behavior, we utilize GPT-4o[11], which we found to have superior accuracy in Named Entity Recognition (NER) and the ability to extract simple plot components from given news articles. Specifically, we use GPT-4o to: 1) extract NEs from news articles, summaries, and sources, and 2) extract plot components including event storyline, event cause and event result from news articles. We then compare the user-annotated sources with the article plot to identify which plot components the user highlights in the summary[12]. To investigate article structures, we project the worker summary source distribution from Section 3.2 as the worker interested position distribution.

4.3 Evaluation results

We report the experimental results of different models using 2-shot prompting in Table 2, where *Generic* refers to generating a summary from the input article without incorporating the user's historical data, with the prompt not accounting for specific factors. *Direct* implies that the prompt does not explicitly include these factors, but it still utilizes the user's previous history. *Plot*, *Entity*, and *Position* indicate that the model prompt is tailored to focus on these particular factors from the user's historical data, with n-shot prompting including this pre-extracted information. The "+" symbol signifies a combination of factors, such as *Entity+Plot*, which considers both. *All* indicates that the prompt instructs the model to account for all relevant factors when generating the summary. We can observe that incorporating the user's historical annotations and personalized factors into the prompt slightly improved the generated results. However, BERTScore struggled to effectively distinguish between different prompt generation results. This may be because most generated summaries convey similar meanings but vary in focus or details. As BERTScore relies on semantic similarity, it loses its advantage in such cases. Besides, through a horizontal comparison of the results generated by

[5]https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

[6]https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-pro?pli=1

[7]https://huggingface.co/NorwAI/NorwAI-Mixtral-8x7B-instruct

[8]We also tested the performance of different models using English prompts, but all results were lower than those using Norwegian prompts.

[9]https://huggingface.co/google-bert/bert-base-uncased

[10]https://huggingface.co/NorGLM/Entailment

[11]https://openai.com/index/hello-gpt-4o

[12]We present our prompts of NER and plot extraction in the supplementary materials.

Table 2: 2-shot experimental results of different LLMs on PersonalSum. Best results are on bold and the second best results are underlined.

Models	Metrics	Generic	Direct	Entity	Plot	Position	Entity+Plot	Entity+Position	Plot+Position	All
GPT-3.5 Turbo	Rouge-1	37.90 ±14.73	38.01 ±14.82	37.56 ±15.23	36.90 ±16.25	37.93 ±15.38	37.93 ±15.36	38.03 ±15.10	<u>38.16</u> ±15.43	**38.43** ±15.22
	Rouge-2	17.00 ±13.04	17.17 ±13.19	16.89 ±13.22	16.55 ±13.52	17.05 ±13.41	17.06 ±13.27	<u>17.27</u> ±13.48	17.20 ±13.71	**17.47** ±13.65
	Rouge-L	26.84 ±13.10	27.16 ±13.13	26.85 ±13.31	26.28 ±14.05	27.15 ±13.65	26.96 ±13.49	<u>27.37</u> ±13.53	27.37 ±13.74	**27.45** ±13.71
	BERTScore	75.00 ±5.39	75.16 ±5.30	74.76 ±5.72	74.64 ±6.14	74.98 ±5.79	75.00 ±5.64	75.02 ±5.62	<u>75.20</u> ±5.64	**75.20** ±5.60
Gemini 1.0 Pro	Rouge-1	35.21 ±13.51	<u>35.67</u> ±14.09	35.30 ±13.46	35.45 ±13.45	35.42 ±13.97	35.60 ±13.89	**35.91** ±14.03	35.62 ±14.05	35.47 ±13.87
	Rouge-2	14.32 ±11.14	<u>14.76</u> ±11.60	14.27 ±11.09	14.37 ±11.02	14.55 ±11.51	14.42 ±11.34	**14.88** ±11.70	14.70 ±11.62	14.59 ±11.43
	Rouge-L	25.21 ±11.86	<u>25.75</u> ±12.52	25.18 ±11.82	25.57 ±11.91	25.46 ±12.22	25.55 ±12.17	**25.86** ±12.38	25.53 ±12.33	25.27 ±11.90
	BERTScore	74.52 ±5.07	**74.74** ±5.24	74.42 ±5.02	74.56 ±5.01	74.49 ±5.28	<u>74.63</u> ±5.14	74.53 ±5.36	74.54 ±5.28	74.41 ±5.28
NorwAI-Mixtral-8x7B-instruct	Rouge-1	33.88 ±12.62	34.14 ±13.54	34.01 ±13.58	33.83 ±13.42	33.96 ±13.31	34.15 ±13.60	33.81 ±13.49	<u>34.24</u> ±13.88	**34.29** ±13.66
	Rouge-2	13.36 ±10.43	13.66 ±11.13	<u>13.77</u> ±11.00	13.56 ±11.13	13.69 ±10.95	13.75 ±11.05	13.62 ±10.96	13.77 ±11.26	**13.89** ±11.03
	Rouge-L	23.58 ±10.49	24.12 ±11.49	24.03 ±11.38	23.99 ±11.55	24.01 ±11.18	**24.16** ±11.51	24.04 ±11.26	24.04 ±11.73	<u>24.13</u> ±11.28
	BERTScore	73.51 ±4.72	73.69 ±4.95	73.79 ±4.94	73.79 ±4.92	<u>73.84</u> ±4.82	73.78 ±5.09	73.75 ±4.94	73.77 ±4.98	**73.95** ±4.87

2/5/10-shot prompting, we found that as the number of user's historical annotations in the prompt increases, performance decreases. The possible reason is that the user's annotation data contains scattered features of interest. For example, if a HIT includes both content the user is interested in and content they are not, the user pays attention to different details and reads the article more deeply for the former, as shown in the sources provided. For the latter, the user tends to read only the beginning or end and provides a general summary.

Through data analysis, we discovered that since articles in each HIT were randomly assigned during data collection, only a small portion of articles within HITs share overlapping entities. As a result, the historical records in the prompt become less relevant to the entity and fail to reflect scenarios where users may focus on specific entities. To better simulate varying user interests in entities, we collected a targeted dataset called Topic-centric PersonalSum, as described in the following section.

5 Topic-centric PersonalSum

Experimental design. For data collection, we initially grouped articles by identical entities. We observed that many extracted NEs were numbers, publishers, and media brands, which were not indicative of user interest. To enhance the possibility that articles contain shared entities other than invalid ones, we set the minimum number of overlapping NEs in articles within the same HIT to 3. Due to budget constraints, we curated 141 HITs following the steps outlined in Section 3.1. Differently, we assigned each HIT to two distinct annotators. Finally, we collected 276 personalized summaries for 72 articles. Among these, 68 articles received at least 2 annotated summaries from distinct annotators, while only 4 articles had a single annotation.

Results and analysis. We conducted the same experiments as described in Section 4 on Topic-centric PersonalSum. Experimental results using 5-shot prompting are shown in Table 3[13]. We can observe the following. First, all personalized results outperform the generic summaries, demonstrating that our data is effectively personalized and captured by different models across various dimensions. Further,

[13]Please see supplementary materials for the statistics and the complete experimental results of the Topic-centric PersonalSum.

https://doi.org/10.52202/079017-3151

Table 3: 5-shot experimental results of different LLMs on Topic-centric PersonalSum. Best results are on bold and the second best results are underlined.

Models	Metrics	Generic	Direct	Entity	Plot	Position	Entity+Plot	Entity+Position	Plot+Position	All
GPT-3.5 Turbo	Rouge-1	37.61 ±13.70	39.75 ±13.86	39.68 ±13.79	**40.22** ±14.55	39.39 ±13.74	39.10 ±14.19	38.68 ±13.60	39.31 ±14.37	39.96 ±13.95
	Rouge-2	16.95 ±11.83	18.13 ±12.44	18.58 ±12.11	18.62 ±12.80	17.90 ±12.20	17.81 ±12.25	17.65 ±11.90	18.06 ±12.85	**18.71** ±12.33
	Rouge-L	26.74 ±12.00	27.86 ±12.34	28.08 ±12.37	28.49 ±13.12	27.99 ±12.46	27.25 ±12.54	27.27 ±11.76	27.91 ±12.91	**28.63** ±12.72
	BERTScore	75.05 ±5.05	75.64 ±5.04	75.57 ±5.23	75.77 ±5.38	75.55 ±5.13	75.60 ±5.06	75.52 ±5.00	75.56 ±5.17	**75.90** ±5.19
Gemini 1.0 Pro	Rouge-1	35.87 ±13.25	37.35 ±13.20	37.54 ±13.49	36.96 ±13.39	37.81 ±13.27	**38.04** ±13.44	36.40 ±13.12	37.50 ±13.56	37.61 ±13.02
	Rouge-2	14.94 ±11.14	15.63 ±10.84	**15.98** ±11.32	15.11 ±10.87	15.95 ±10.99	15.91 ±11.20	14.85 ±10.73	15.57 ±11.42	15.74 ±10.98
	Rouge-L	25.05 ±11.82	25.96 ±11.68	**26.47** ±12.30	25.48 ±11.86	26.09 ±11.70	26.45 ±12.36	25.22 ±11.24	26.06 ±12.00	25.94 ±11.38
	BERTScore	74.50 ±5.10	75.02 ±5.13	75.03 ±5.03	74.98 ±5.07	74.92 ±5.00	**75.29** ±5.24	74.70 ±4.99	75.03 ±5.08	74.96 ±5.01
NorwAI-Mixtral-8x7B-instruct	Rouge-1	33.40 ±12.17	35.29 ±13.87	33.83 ±13.80	34.28 ±13.78	34.67 ±13.24	34.83 ±14.03	34.36 ±13.98	**35.75** ±13.91	34.60 ±14.08
	Rouge-2	13.05 ±9.57	14.49 ±11.29	13.56 ±10.50	13.86 ±10.67	13.82 ±10.69	14.51 ±11.10	14.18 ±10.76	**14.98** ±11.01	14.23 ±11.00
	Rouge-L	23.04 ±10.04	24.59 ±11.50	23.60 ±10.96	24.24 ±11.48	23.93 ±11.26	24.82 ±11.94	24.39 ±11.55	**24.99** ±11.70	24.58 ±11.95
	BERTScore	73.41 ±4.60	74.21 ±4.96	73.78 ±5.12	73.86 ±5.15	73.96 ±4.98	74.27 ±5.15	74.16 ±5.11	**74.38** ±5.18	74.12 ±5.19

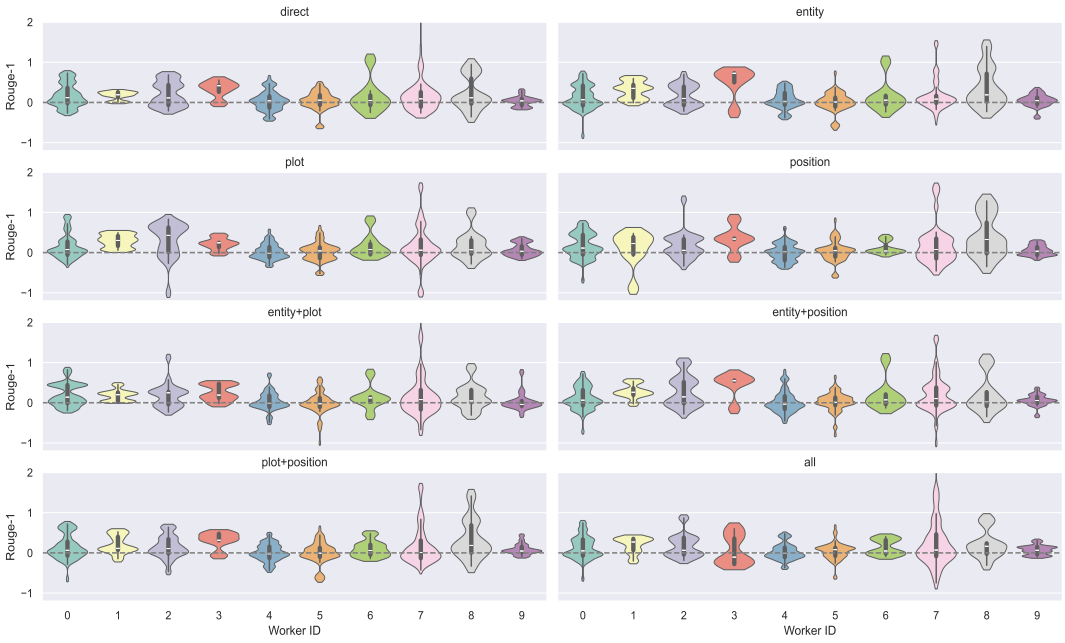

Figure 3: Experimental results showing improvements in the ROUGE-1 score from personalized prompting compared to generic summaries using GPT-3.5 Turbo for each worker. The X-axis represents worker IDs, and the Y-axis represents the ROUGE-1 score improvements.

https://doi.org/10.52202/079017-3151

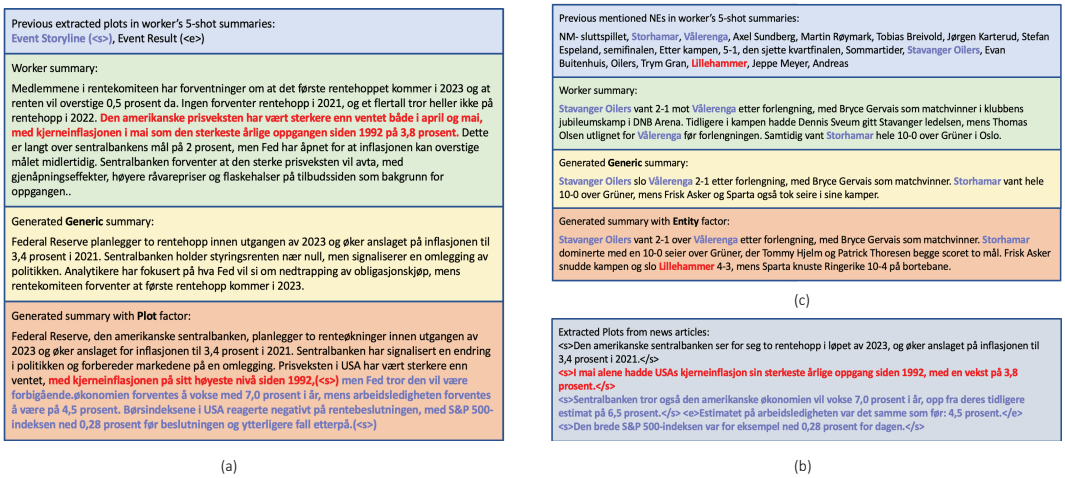

Figure 4: (a) The plot information concerned in the 5-shot historical annotated summaries of Worker 3, the generic summary, and the summary with the prompt including the annotator's plot information. (b) The article's plot data is extracted by GPT-4o. For clarity, we only include the original information relevant to the generated summaries for Worker 3. (c) The entities that appear in the 5-shot historical annotations of Worker 1, the user-annotated summary, the generic summary, and the summary with the prompt including the annotator's entity details. All generated summaries are from GPT-3.5-Turbo.

explicitly incorporating diverse factors into the prompt influences the model's output to varying extents. We also observe that 5-shot prompting yields the best results across all models, whereas 10-shot prompting performs worse than 2-shot prompting. This indicates that when generating personalized summaries, it is crucial to balance the number of input user history records. Compared with Section 4.3, although the best results appear with different few-shot prompting methods, we can still see that an excessive amount of user history data introduces noise to the pre-trained models, adversely affecting the generation outcomes. The superior experimental results on Topic-centric PersonalSum may demonstrate that it exhibits more pronounced user annotation characteristics compared to PersonalSum.

To gain a more comprehensive understanding of how different factors impact the models, we investigated the performance from the user side. Figure 3 shows the experimental results of the improvements in the ROUGE-1 score from personalized prompting compared to generic summaries using GPT-3.5 Turbo for each worker, based on Table 3. From the differences in improvement, we selected two instances for further analysis: one from worker 1, who showed a higher improvement, and one from worker 3, who showed reduced performance. This analysis aims to have a hint on the impact of various factors. As shown in Figure 4(a) and 4(b), Worker 3 is interested in event storylines which could be details of the event, and event results. When the model is prompted to generate a summary with a storyline, it introduces additional descriptions about 1992 that align with the user's annotations, compared to the generic summary. However, it also includes descriptions of the S&P 500 index, which were not annotated by the user. In Figure 4(c), when the model is prompted to generate a summary considering Worker 1's previously annotated NEs, apart from the entities highlighted in purple that match the annotated summary, the model also includes the entity "Lillehammer" which appears in the user's history.

6 Human Evaluation

We recruited three well-educated Norwegian native colleague students to evaluate model generated summaries (including generic, direct and all factors) of three models (GPT3.5-turbo, Gemini, and NorwAI-Mixtral-8x7B-instruct) using 5-shot prompting with human written summaries. Considering the time and cost limit, we randomly selected 50 samples from PersonalSum for evaluation. The detailed instructions to the evaluators are shown in Figure 10 in the supplementary materials.

We adopted Fleiss' kappa (κ) to measure Inter-rater Agreement among the three raters for each evaluation metric and model. The results are shown in Table 4. The Fleiss' kappa score shows that all

https://doi.org/10.52202/079017-3151

Table 4: Human evaluation results on the quality of personalized summaries generated by LLMs.

Models	Consistency / Fleiss' kappa			Coherence / Fleiss' kappa		
	Generic	Direct	All	Generic	Direct	All
GPT-3.5 Turbo	4.03 / 0.96	4.02 / 0.91	**4.05** / 0.91	**4.78** / 0.86	4.77 / 0.80	4.70 / 0.86
Gemini 1.0 Pro	3.95 / 0.83	4.01 / 0.73	**4.03** / 0.86	4.69 / 0.82	**4.74** / 0.82	4.67 / 0.98
NorwAI-Mixtral-8x7B-instruct	3.81 / 0.82	3.87 / 0.71	**3.99** / 0.83	4.53 / 0.66	4.59 / 0.83	**4.63** / 0.77

evaluation results achieve substantial or almost perfect agreement [35][14]. From the Table, we can see that while the generic summary preserves much of the content from the user-annotated summary, the summary generated using prompts that explicitly include entities, news plots, or news article structure preference aligns more closely with the user's personalized content needs. In addition, after analyzing the issues present in the generated summaries provided by the raters, we observed that for GPT3.5-Turbo and Gemini1.0-pro, the primary challenges are "2. excessive detail", followed by "1. a focus on different topics" and "4. divergent plot emphasis". In contrast, the primary issues for NorwAI-Mixtral-8x7B-instruct involve "1. a focus on different topics", followed by "4. divergent plot emphasis" and "5. incomplete outputs".

7 Concluding remarks

From the experimental results on both datasets, we observe the following: 1) Entities play a crucial role in personalized summarization. Despite the similarity or interconnectedness of topics in many HITs articles in the Topic-centric dataset, solely considering the entity factor in few-shot scenarios may not maximize improvement. Plot and article structure could also be among the myriad factors affecting the user's personalized summary. 2) Through the analysis of individual use cases, we found that there are many details in the user's personalized summary, which are often ignored by expert writers and journalists when writing summaries. This is reasonable because professional writers and journalists often extract the main points and salient content to meet the needs of the public. However, we argue that the uniqueness of each individual should not be ignored. (3) Limitations of our work include an insufficient amount of data for model training and the workers were not able to select articles to annotate themselves.

The rich properties of PersonalSum enable its use across various applications, such as evaluating the explainability and factuality of document-grounded question-answering systems and news summarization models, exploring information extraction characteristics, and uncovering the general public's implicit interests. Furthermore, it provides an opportunity to delve into a comparative analysis of machine-generated summaries and those annotated manually. This diverse range of applications underscores the versatility and potential of the dataset in advancing research in these areas.

8 Ethical Statement

Prior to the annotation process, all participants were informed about the purpose and intended use of the data they provided. Annotators were given the option to select "prefer not to say" to ensure they feel comfortable sharing personal information. Sensitive data, such as age and occupation, was used solely for statistical purposes and will not be shared or used beyond the scope of the study outlined in this paper. This ensures that participants' identities remain unidentifiable through the annotated data. Also, each users were given all necessary information about the extents of our study, hence ensuring transparency.

To achieve a diverse representation of users, we collected annotations from various age groups, including those under 25 and over 35, since the majority of annotators were aged 25-34. While biases may arise if users' preferences do not align with the assigned news articles, capturing user subjectivity is one of our goals, as it reflects real user behavior and personal perspectives.

We emphasize that this work focuses on exploring the capability of LLMs to generate personalized summaries rather than defining users by their traits or personalities, thus respecting privacy and avoiding unwarranted generalizations. The code and dataset used for data collection and experiments in this paper have been made publicly available on GitHub for reproducibility purposes.

[14]https://www.ncbi.nlm.nih.gov/books/NBK92287/table/executivesummary.t2/?report=objectonly

Acknowledgments

This work was carried out at the Norwegian Research Center for AI Innovation (NorwAI), funded by the Research Council of Norway through the Centre for Research-based Innovation (SFI) funding scheme, with additional financial support from NorwAI's partners.

We extend our gratitude to the reviewers for their valuable feedback. Special thanks to the IDUN team at NTNU [36] for providing essential computational resources, and to Schibsted and the National Library of Norway (Nasjonalbiblioteket) for supplying the crucial dataset for our research.

References

[1] Kukoleva Olesya, Anna Preobrazhenskaya, and Olga Sidorova. 2017. Media use habits: what, why, when, and how people read online. UXMatters. https://www.uxmatters.com/mt/archives/2017/07/media-use-habits-what-why-when-and-how-people-read-online.php.

[2] Hermann, Karl Moritz, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems (NIPS), pages 1684–1692.

[3] Narayan, Shashi, Shay B. Cohen, and Mirella Lapata. 2018. 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, pp. 1797-1807.

[4] Grusky, Max, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 708-719.

[5] Sharma, Eva, Chen Li, and Lu Wang. 2019. BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2204-2213.

[6] Cohan, Arman, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 615-621.

[7] Hu, Baotian, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A Large Scale Chinese Short Text Summarization Dataset. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1967-1972.

[8] Koupaee, Mahnaz, and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305.

[9] Dang, Hoa Trang, Lucy Vanderwende, Catherine Blake, Julia Kampov, Andreas Orphanides, David West, Cory Lown, Massih Amini, Nicolas Usunier, and Fabrizio Gotti. 2007. Document understanding conference duc 2007. In Document Understanding Conference.

[10] Zhong, Ming, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan et al. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5905-5921.

[11] Hayashi, Hiroaki, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. Wikiasp: A dataset for multi-domain aspect-based summarization. Transactions of the Association for Computational Linguistics 9 (2021): 211-225.

 https://doi.org/10.52202/079017-3151

[12] Zhang, Yusen, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023. Macsum: Controllable summarization with mixed attributes. Transactions of the Association for Computational Linguistics 11 (2023): 787-803.

[13] McAuley, Julian, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pp. 43-52.

[14] Ao, Xiang, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 82-92.

[15] Xu, Hongyan, Hongtao Liu, Zhepeng Lv, Qing Yang, and Wenjun Wang. 2023. Pre-trained Personalized Review Summarization with Effective Salience Estimation. In Findings of the Association for Computational Linguistics: ACL 2023, pp. 10743-10754.

[16] Cheng, Xin, Shen Gao, Yuchi Zhang, Yongliang Wang, Xiuying Chen, Mingzhe Li, Dongyan Zhao, and Rui Yan. 2023. Towards personalized review summarization by modeling historical reviews from customer and product separately. arXiv preprint arXiv:2301.11682.

[17] Xu, Hongyan, Hongtao Liu, Zhepeng Lv, Qing Yang, and Wenjun Wang. 2023. Sentiment-aware Review Summarization with Personalized Multi-task Fine-tuning. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pp. 2826-2835.

[18] Wysoczanska, Monika, Moran Beladev, Karen Lastmann Assaraf, Fengjun Wang, Ofri Kleinfeld, Gil Amsalem, and Hadas Harush Boke. 2024. Tell Me What Is Good About This Property: Leveraging Reviews For Segment-Personalized Image Collection Summarization. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 21, pp. 22983-22989.

[19] Dharan, Nidhin S., and R. Gowtham. 2021. Personalized Abstract Review Summarization Using Personalized Key Information-Guided Network. In Inventive Computation and Information Technologies: Proceedings of ICICIT 2021, pp. 203-216.

[20] Li, Junjie, Haoran Li, and Chengqing Zong. 2019. Towards personalized review summarization via user-aware sequence network. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 6690-6697.

[21] Xu, Hongyan, Hongtao Liu, Pengfei Jiao, and Wenjun Wang. 2021. Transformer reasoning network for personalized review summarization. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1452-1461.

[22] Chan, Hou Pong, Wang Chen, and Irwin King. 2020. A unified dual-view model for review summarization and sentiment classification with inconsistency loss. In Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, pp. 1191-1200.

[23] Li, Piji, Zihao Wang, Lidong Bing, and Wai Lam. 2019. Persona-aware tips generation?. In The World Wide Web Conference, pp. 1006-1016.

[24] Liu, Hui, and Xiaojun Wan. 2019. Neural review summarization leveraging user and product information. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2389-2392.

[25] Urlana, Ashok, Pruthwik Mishra, Tathagato Roy, and Rahul Mishra. 2023. Controllable Text Summarization: Unraveling Challenges, Approaches, and Prospects–A Survey. arXiv preprint arXiv:2311.09212.

[26] He, Junxian, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards Generic Controllable Text Summarization. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 5879-5915.

[27] Liu, Yixin, Alexander R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In Findings of the Association for Computational Linguistics: NAACL 2024, pp. 4481-4501.

[28] Dou, Zi-Yi, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A General Framework for Guided Neural Abstractive Summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4830-4842.

[29] Wu, Jeff, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. arXiv preprint arXiv:2109.10862.

[30] Lin, Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pp. 74-81.

[31] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In International Conference on Learning Representations. 2019.

[32] Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1906-1919.

[33] Liu, Peng, Lemei Zhang, Terje Nissen Farup, Even W. Lauvrak, Jon Espen Ingvaldsen, Simen Eide, Jon Atle Gulla, and Zhirong Yang. 2023. NLEBench + NorGLM: A Comprehensive Empirical Analysis and Benchmark Dataset for Generative Language Models in Norwegian. arXiv preprint arXiv:2312.01314.

[34] Kutuzov, Andrey, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-Scale Contextualised Language Modelling for Norwegian. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 30–40.

[35] Landis, J.R. and Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics, pp.159-174.

[36] Själander, Magnus, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. 2019. EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure. arXiv preprint arXiv:1912.05848.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

 https://doi.org/10.52202/079017-3151

(b) Did you describe the limitations of your work? [Yes] See Section 4.3 and Section 7

(c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 8

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [N/A]

(b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and Section 4

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5 and Section 4

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5 and Section 4

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4

(b) Did you mention the license of the assets? [Yes] See Section 1 and Section 4

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See supplemental materials

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Section 4