

NEURAL INFORMATION
PROCESSING SYSTEMS

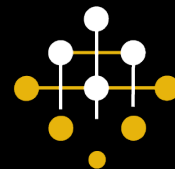
PersonalSum: A User-Subjective Guided Personalized Summarization Dataset for Large Language Models

Lemei Zhang, Peng Liu, Marcus Tiedemann Oekland Henriksboe,
Even W. Lauvrak, Jon Atle Gulla, Heri Ramampiaro

December 2024

NorwAI

Norwegian Research Center
for AI Innovation



NTNU

sfi Centre for
Research-based
Innovation

The Research Council of Norway

Background and Motivation

- Three distinct news article reading habits of users
 - **Attentive reading:** users read the full article attentively, focusing on details
 - **Selective reading:** users focus only on interesting fragments
 - **Scanning:** users absorb only the important ideas



Kukoleva Olesya, Anna Preobrazhenskaya, and Olga Sidorova. 2017. *Media use habits: what, why, when, and how people read online*. UXMatters.

Background and Motivation

- LLMs demonstrate remarkable proficiency in generating **high-quality generic summaries**, even surpassing those produced by experts, according to human evaluations.
- Challenge: ***Are users interested in the information presented in a generic summary?***

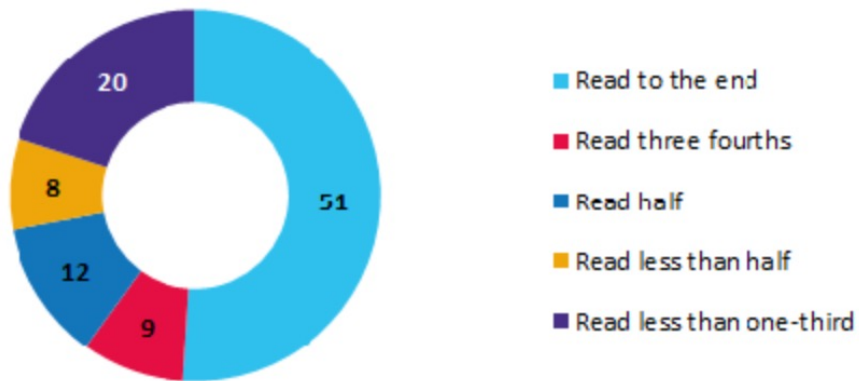
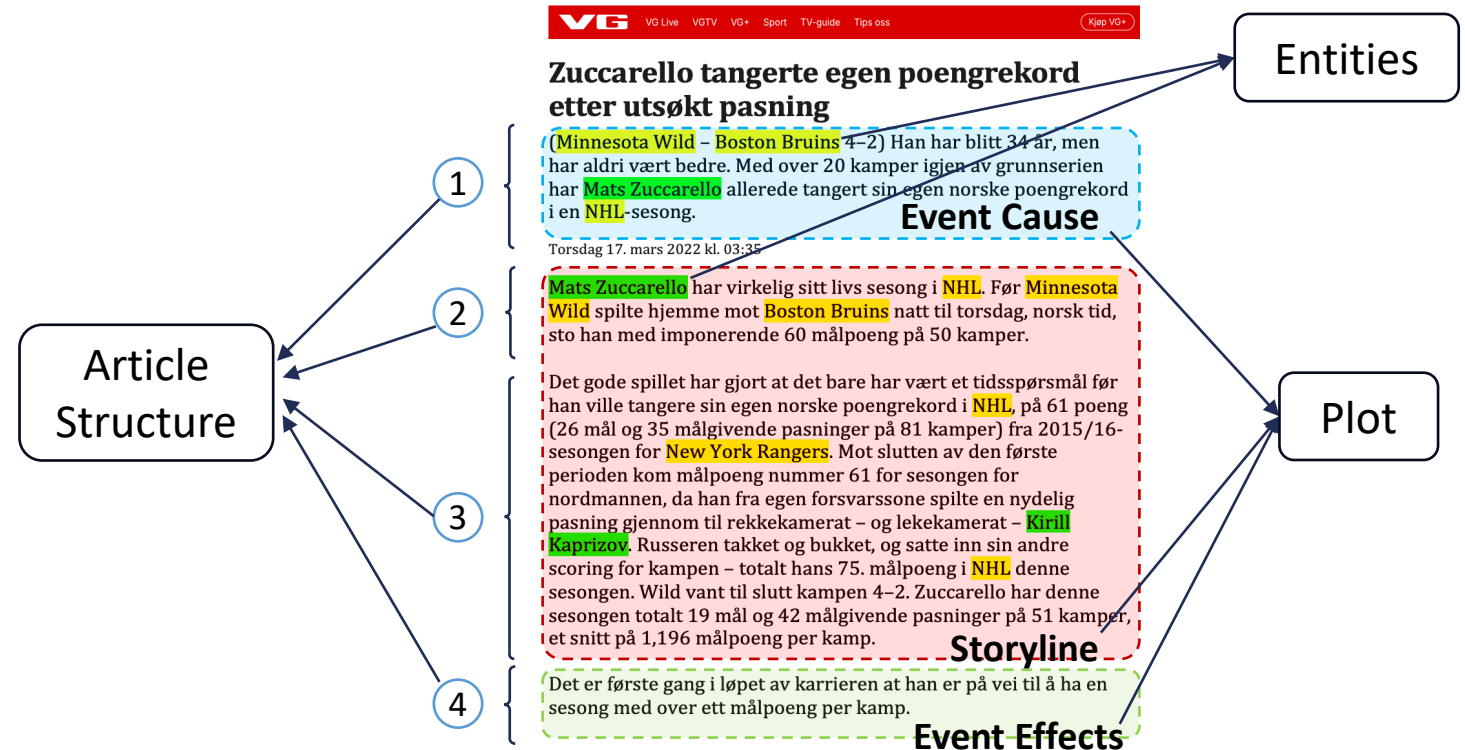


Figure: News reading depth on the desktop.
Image credit to *Kukoleva et al. 2017*



Background and Motivation

- Existing work on **personalized summarization** often relies on pseudo datasets created from **generic summarization datasets** or **controllable datasets** 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.

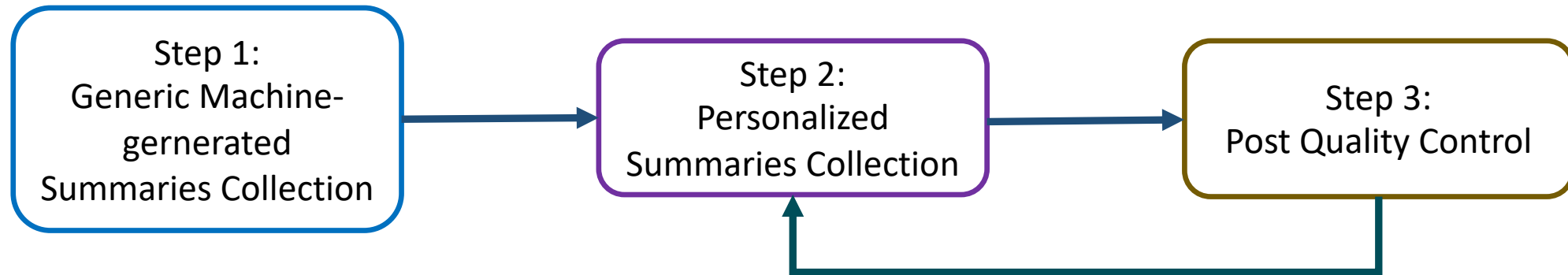
Table 1: Comparison between PersonalSum and existing popular summarization datasets.

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	✓	✓	✓	✓	✓



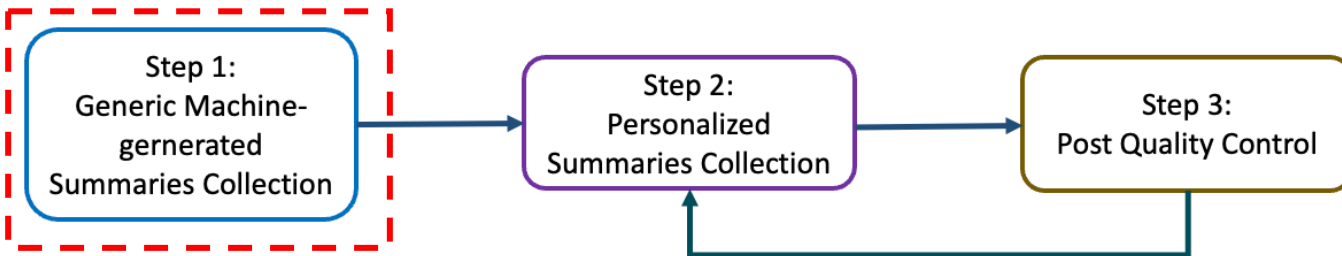
PersonalSum: A User-Subjective Guided Personalized Summarization Dataset

Data Collection



Data Collection

- **Step 1:** Collection of Machine-generated Generic Summaries
 - Some statistics: 465 news articles, 10 categories
 - Generated summaries: GPT-4
 - 3 Norwegian students for quality control: 100% inter-agreement
 - Annotated attributes: summaries, document-grounded question answers, sources

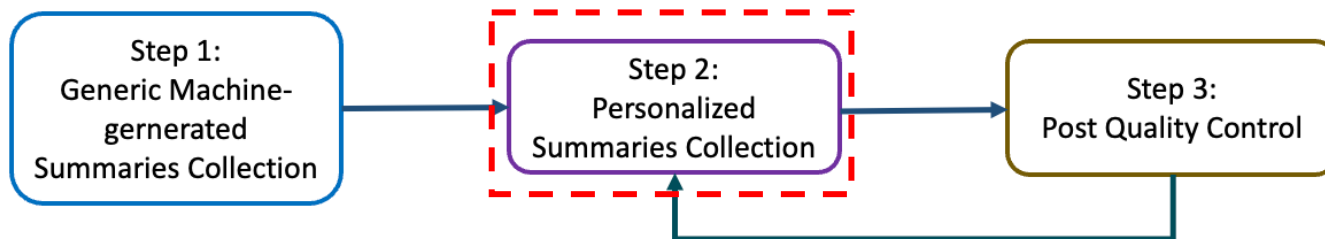


The screenshot shows a web application interface. At the top, there is a title: "Johannes Høstflot Klæbo kjøpte nabohuset for 13,3 millioner kroner". Below the title, there are fields for "Newaroom: ap" and "Creation_date: 2022-09-16T14:00:14Z". The main content area displays a news article body. Below the article, there is a "Questions & Answers" section with a "Summary" tab. The "Question" field contains: "Hvor mye betalte Johannes Høstflot Klæbo for nabohuset sitt, og hva var opprinnelig prisen?". The "Answer" field contains: "Johannes Klæbo betalte 13,3 millioner kroner, og den opprinnelige prisen ble satt til 11,2 millioner kroner." The "Source" field contains: "Johannes Høstflot Klæbos nabohus lå ute til salg for 11,2 millioner. Det ble solgt for 13,3 millioner." There are buttons for "Download GPT-Questions", "Download GPT-Summary", "Generate Question", "ADD", and "Save to file".



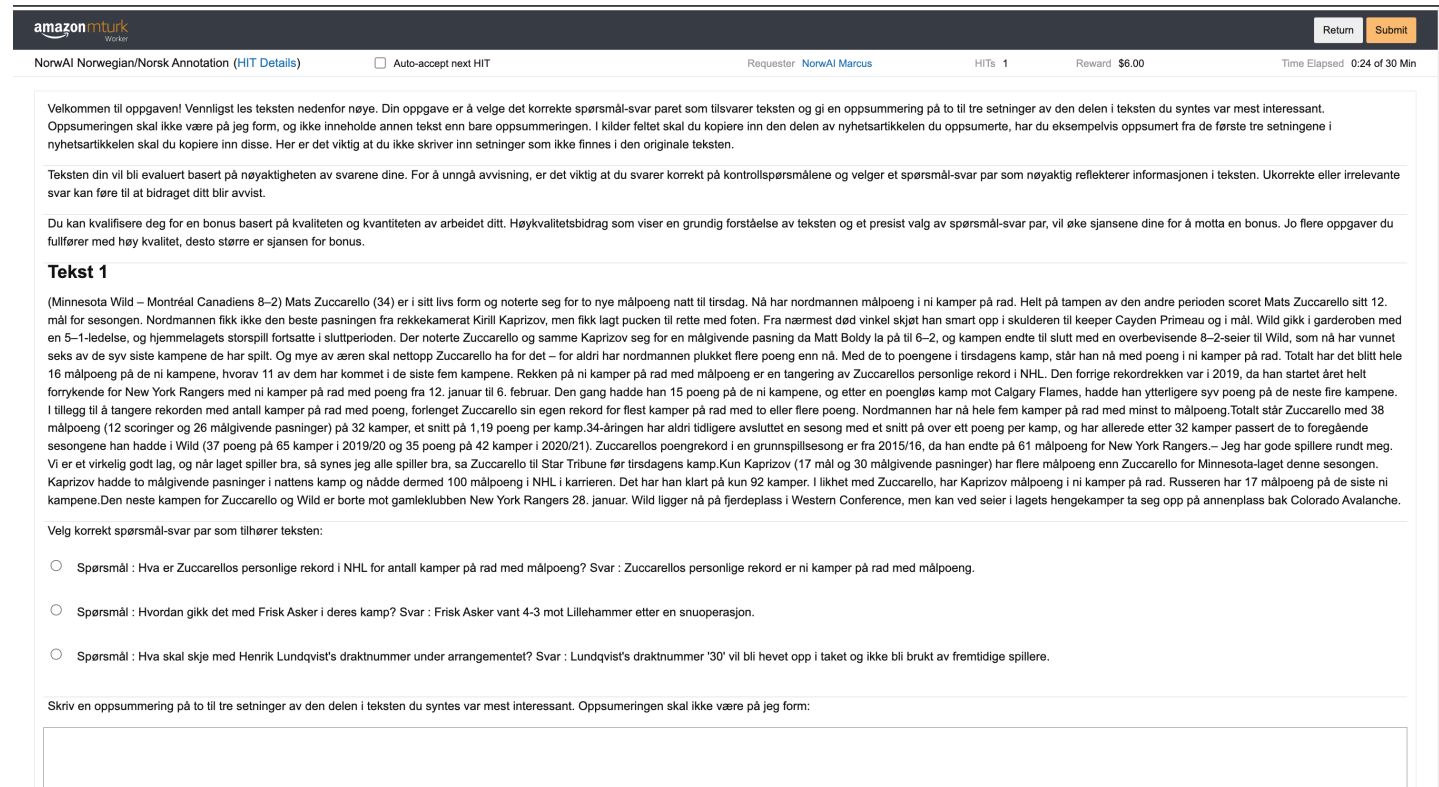
Data Collection

- **Step 2:** Collection of Personalized Summaries
 - Crowdsourcing platform: Amazon Mechanical Turk
 - Questionnaire of Qualification Test before assignment: fluency in Norwegian, demographic information, news consumption habits, areas of interest, and gender
 - 3 articles per HIT; One HIT is assigned to 3 annotators;
 - Qualification Test for annotation: 3 single-choice questions about the articles / HIT
 - Accuracy rate: $>2/3$



Data Collection

- **Step 2:** Collection of Personalized Summaries
 - Automatic filtering rules:
 - Summary length: >50 words
 - Task duration: >5 minutes



The screenshot shows an Amazon MTurk HIT page. At the top, it displays the Amazon MTurk logo and the requester's name, NorwAI Marcus. The HIT title is "NorwAI Norwegian/Norsk Annotation (HIT Details)". The page includes a "Return" button and a "Submit" button. The main content area contains instructions in Norwegian, a text sample, and a list of three multiple-choice questions. The text sample discusses the performance of Mats Zuccarello and Kirill Kaprizov in the NHL. The questions are:

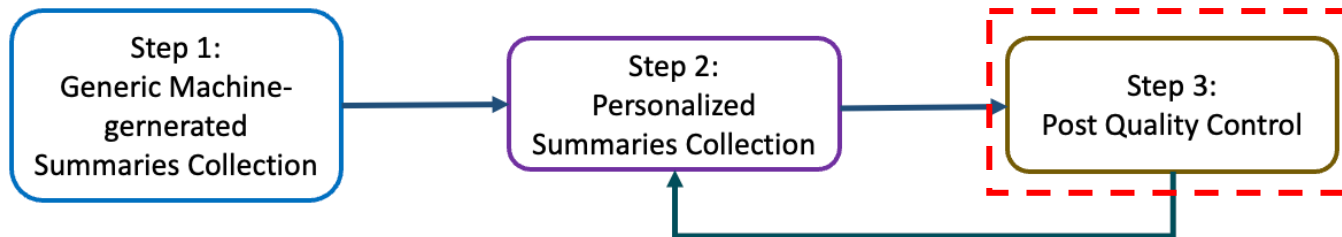
- Spørsmål : Hva er Zuccarellos personlige rekord i NHL for antall kamper på rad med målpoeng? Svar : Zuccarellos personlige rekord er ni kamper på rad med målpoeng.
- Spørsmål : Hvordan gikk det med Frisk Asker i deres kamp? Svar : Frisk Asker vant 4-3 mot Lillehammer etter en snuoperasjon.
- Spørsmål : Hva skal skje med Henrik Lundqvist's draktnummer under arrangementet? Svar : Lundqvist's draktnummer '30' vil bli hevet opp i taket og ikke bli brukt av fremtidige spillere.

Below the questions, there is a text box for the user to provide a summary of the text.

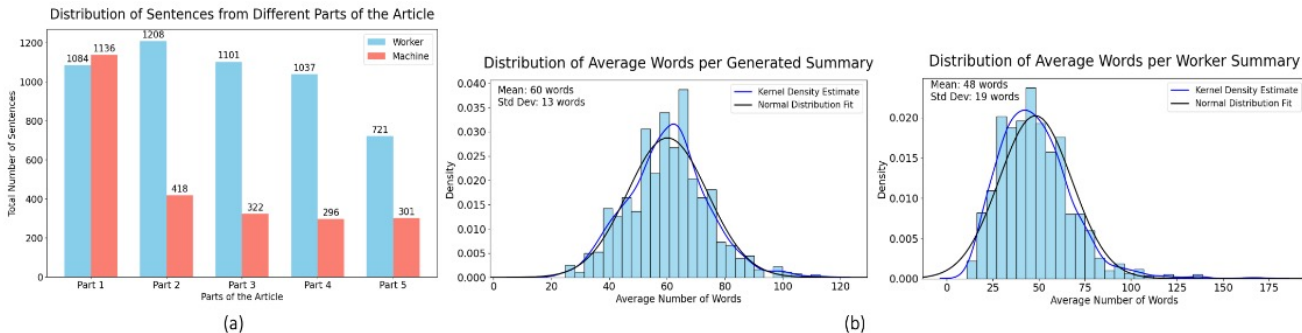
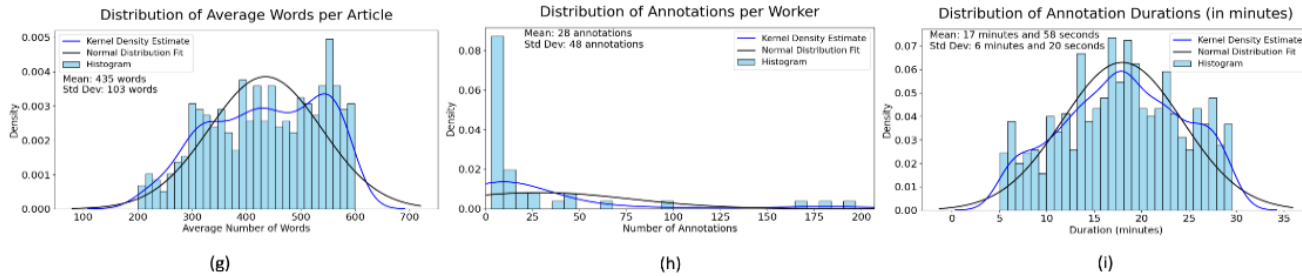
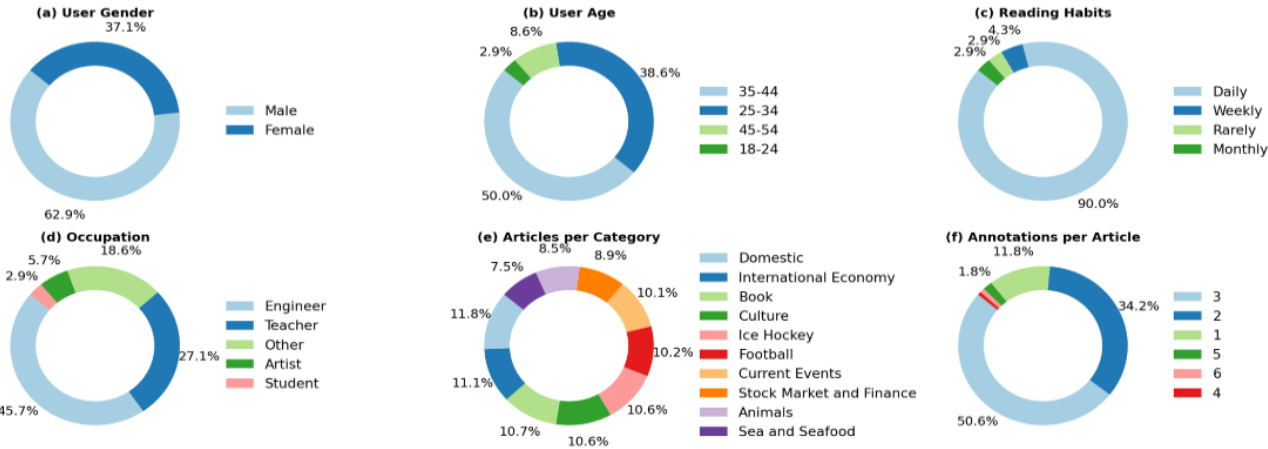


Data Collection

- **Step 3: Post Quality Control**
 - LLM evaluation:
 - Evaluate the annotated summary to a given article as well as its sources
 - Metrics: Coherence, Consistency, and Relevance
 - Human evaluation:
 - 10% of random sampled data for relevance score > 0.8
 - All annotations for relevance score ≤ 0.8



Data Statistics



- Machine-generated summaries are relatively longer than human-annotated ones;
- Most machine-generated summaries originate from the first part of the article;
- Human annotated summaries are relatively evenly distributed across various parts of the article;
- 91% of annotations for the same article are from different parts of the article.



Experiments

- Experimental settings:
 - Base models: OpenAI GPT-3.5 Turbo, Llama3-instruct, Google Gemini-1.0-pro, NorwAI-Mixtral-8x7B instruct;
 - Methods: zero-shot prompting, 1/5/10-shot prompting;
 - Impact factors: Named Entities (NE), article plot, news structure
 - Objectives: If pretrained models can capture personalized signals



Experimental Results

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	<u>75.02</u> ±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	<u>35.47</u> ±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	<u>33.81</u> ±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

- Columns:

- Generic: no user's historical data
- Direct: keep user's historical data for few-shot prompt, but no explicit factor
- Other columns represent the model prompt is tailored to focus on particular factor(s) from the user's historical data

- As the number of user's historical annotations in the prompt increases, performance decreases. Possibly due to the scattered features of interest in users' history



Experimental Results

Few overlapped NEs in one HIT!

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	<u>35.47</u> ±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	<u>33.81</u> ±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

- Columns:

- Generic: no user's historical data
- Direct: keep user's historical data for few-shot prompt, but no explicit factor
- Other columns represent the model prompt is tailored to focus on particular factor(s) from the user's historical data

- As the number of user's historical annotations in the prompt increases, performance decreases. Possibly due to the scattered features of interest in users' history



Topic-Centric PersonalSum

- Data collection:
 - Grouped articles by **identical NEs**
 - Follow the same collection process as previous dataset
 - 72 articles, 276 personalized summaries



Experimental Results- Topic-Centric PersonalSum

Table 3: 2-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	37.14 ±14.02	39.16 ±14.13	39.21 ±14.36	<u>39.42</u> ±15.09	39.37 ±14.18	38.78 ±14.36	39.90 ±14.42	39.30 ±14.01	39.07 ±14.46
	Rouge-2	16.43 ±12.06	17.58 ±12.30	17.94 ±12.78	18.15 ±12.84	17.94 ±12.42	17.86 ±12.43	18.81 ±12.75	<u>18.22</u> ±12.33	17.61 ±12.35
	Rouge-L	26.43 ±12.30	27.23 ±12.20	27.72 ±12.86	28.03 ±13.26	27.70 ±12.40	27.27 ±12.37	<u>28.06</u> ±12.56	28.12 ±12.67	27.50 ±12.54
	BERTScore	74.84 ±5.13	75.57 ±5.20	75.60 ±5.38	75.59 ±5.59	75.66 ±5.25	75.51 ±5.30	75.75 ±5.37	<u>75.73</u> ±5.13	75.54 ±5.32
Gemini 1.0 Pro	Rouge-1	35.46 ±12.84	36.48 ±12.96	36.30 ±13.25	35.26 ±14.32	35.14 ±14.53	<u>36.46</u> ±14.10	<u>36.32</u> ±13.54	36.18 ±13.86	36.33 ±13.81
	Rouge-2	14.52 ±11.02	14.74 ±10.41	15.04 ±10.35	13.89 ±10.46	14.24 ±10.93	15.19 ±11.26	14.99 ±11.38	<u>15.11</u> ±11.22	15.06 ±11.11
	Rouge-L	24.88 ±11.16	25.74 ±11.59	25.73 ±11.17	24.54 ±12.20	24.63 ±11.95	25.88 ±12.56	25.75 ±12.17	25.59 ±12.42	<u>25.76</u> ±11.89
	BERTScore	74.45 ±4.95	<u>74.82</u> ±4.98	74.66 ±5.20	74.32 ±5.61	73.85 ±6.41	74.85 ±5.49	74.72 ±5.44	74.60 ±5.54	74.76 ±5.11
NorwAI-Mixtral-8x7B-instruct	Rouge-1	32.80 ±11.98	<u>33.87</u> ±12.14	33.20 ±11.85	33.23 ±13.19	33.65 ±12.49	<u>33.19</u> ±12.57	34.32 ±12.90	33.58 ±12.71	33.34 ±13.17
	Rouge-2	12.29 ±9.47	13.28 ±9.55	12.61 ±8.87	13.07 ±9.92	13.04 ±9.65	12.92 ±9.68	13.82 ±10.22	12.66 ±9.31	<u>13.35</u> ±10.21
	Rouge-L	22.66 ±10.26	<u>23.62</u> ±10.26	22.68 ±9.24	23.01 ±10.78	23.45 ±10.17	22.96 ±10.18	23.93 ±10.97	23.23 ±10.14	23.32 ±10.82
	BERTScore	73.15 ±4.57	<u>73.77</u> ±4.52	73.63 ±4.41	73.73 ±4.75	73.60 ±4.51	73.69 ±4.72	74.13 ±4.80	73.60 ±4.72	73.72 ±4.80
Meta-Llama3-70B-Instruct	Rouge-1	35.98 ±13.02	17.86 ±14.97	16.73 ±13.82	17.88 ±16.50	18.30 ±16.23	17.86 ±15.96	16.60 ±14.99	18.61 ±16.45	18.05 ±15.80
	Rouge-2	14.15 ±10.29	7.20 ±8.88	6.74 ±8.00	7.33 ±9.36	7.54 ±9.61	7.43 ±8.98	6.48 ±8.24	7.66 ±9.41	7.29 ±9.30
	Rouge-L	24.25 ±10.21	13.56 ±10.86	12.87 ±9.97	13.17 ±11.67	13.49 ±11.60	13.43 ±11.35	12.52 ±10.95	13.69 ±11.63	13.33 ±11.28
	BERTScore	74.10 ±4.99	71.03 ±4.62	69.93 ±5.09	70.05 ±6.50	70.48 ±5.75	70.31 ±6.11	69.78 ±6.10	70.59 ±5.64	70.03 ±6.10

- all personalized results outperform the generic summaries;
- Explicitly incorporating diverse personalized signals into the prompt affects the model's output to varying extents;
- Performance of 10-shot prompting worse than 5-shot and 2-shot prompting



Human Evaluation

- 50 samples on each factors from 5-shot results of PersonalSum:
 - Investigation factors: generic, direct and all factors
 - Models: GPT3.5-turbo, Gemini, and NorwAI-Mixtral-8x7B-instruct
- Evaluation metrics: Consistency and Coherence
- Internal agreement: Fleiss' kappa (κ)

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



Human Evaluation

- Results:
 - Explicitly incorporate personalized factors performs better than *Generic* and *Direct*
 - Problem with GPT3.5-Turbo and Gemini 1.0-pro: *Excessive Details, Focus on Different Topics, Divergent Plot Emphasis*
 - Problem with NorwAI-Mixtral-8x7B-instruct: *Focus on Different Topics, Divergent Plot Emphasis, Incomplete Output*

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



Case Study

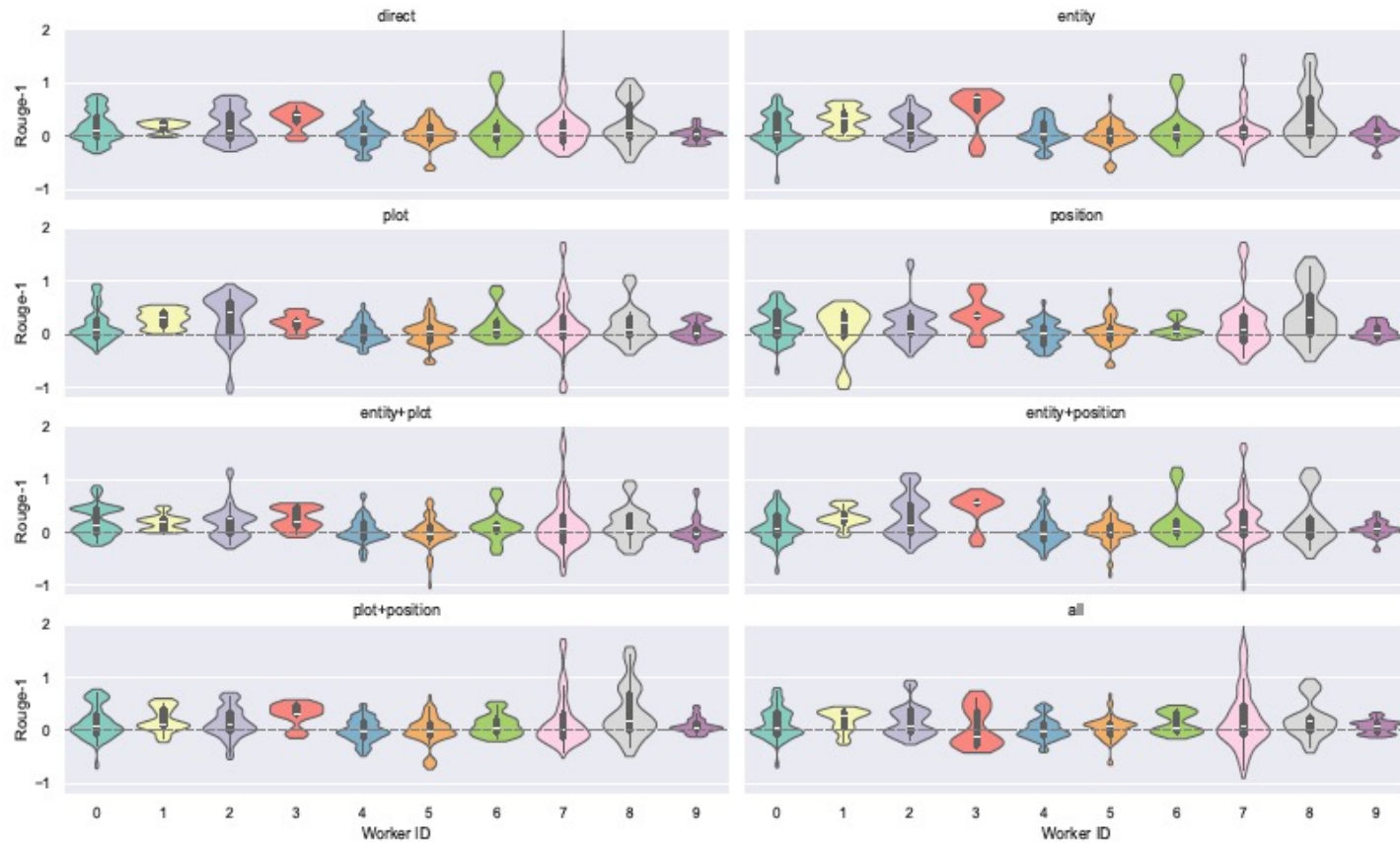


Figure 3: Experimental results showing improvements in the ROUGE-1 score from personalized prompting compared to generic summaries using GPT-3.5 Turbo for each worker. The X-axis represents worker IDs, and the Y-axis represents the ROUGE-1 score improvements.

- Two instances for further analysis w.r.t. ROUGE1 score on 5-shot prompting:
 - Worker 1: higher improvements
 - Worker 3: reduced improvements

Case Study

<p>Previous extracted plots in worker's 5-shot summaries: Event Storyline (<=>), Event Result (<=>)</p>
<p>Worker summary:</p> <p>Medlemmene i rentekomiteen har forventninger om at det første rentehoppet kommer i 2023 og at renten vil overstige 0,5 prosent da. Ingen forventer rentehopp i 2021, og et flertall tror heller ikke på rentehopp i 2022. Den amerikanske prisveksten har vært sterkere enn ventet både i april og mai, med kjerneinflasjonen i mai som den sterkeste årlige oppgangen siden 1992 på 3,8 prosent. Dette er langt over sentralbankens mål på 2 prosent, men Fed har åpnet for at inflasjonen kan overstige målet midlertidig. Sentralbanken forventer at den sterke prisveksten vil avta, med gjenåpningseffekter, høyere råvarepriser og flaskehalsar på tilbudssiden som bakgrunn for oppgangen...</p>
<p>Generated Generic summary:</p> <p>Federal Reserve planlegger to rentehopp innen utgangen av 2023 og øker anslaget på inflasjonen til 3,4 prosent i 2021. Sentralbanken holder styringsrenten nær null, men signaliserer en omlegging av politikken. Analytikere har fokusert på hva Fed vil si om nedtrapping av obligasjonskjøp, mens rentekomiteen forventer at første rentehopp kommer i 2023.</p>
<p>Generated summary with Plot factor:</p> <p>Federal Reserve, den amerikanske sentralbanken, planlegger to rentetøkninger innen utgangen av 2023 og øker anslaget for inflasjonen til 3,4 prosent i 2021. Sentralbanken har signalisert en endring i politikken og forbereder markedene på en omlegging. Prisveksten i USA har vært sterkere enn ventet, med kjerneinflasjonen på sitt høyeste nivå siden 1992,(<=>) men Fed tror den vil være forbigående. Økonomien forventes å vokse med 7,0 prosent i år, mens arbeidsledigheten forventes å være på 4,5 prosent. Børsindeksene i USA reagerte negativt på rentebeslutningen, med S&P 500-indeksen ned 0,28 prosent før beslutningen og ytterligere fall etterpå.<=></p>

(a)

<p>Previous mentioned NEs in worker's 5-shot summaries: NM-sluttspillet, Storhamar, Vålerenga, Axel Sundberg, Martin Røymark, Tobias Breivold, Jørgen Karterud, Stefan Espeland, semifinalen, Etter kampen, 5-1, den sjette kvartfinalen, Sommertider, Stavanger Oilers, Evan Buitenhuis, Oilers, Trym Gran, Lillehammer, Jeppe Meyer, Andreas</p>
<p>Worker summary:</p> <p>Stavanger Oilers vant 2-1 mot Vålerenga etter forlengning, med Bryce Gervais som matchvinner i klubbens jubileumskamp i DNB Arena. Tidligere i kampen hadde Dennis Sveum gitt Stavanger ledelsen, mens Thomas Olsen utlignet for Vålerenga før forlengningen. Samtidig vant Storhamar hele 10-0 over Grüner i Oslo.</p>
<p>Generated Generic summary:</p> <p>Stavanger Oilers slo Vålerenga 2-1 etter forlengning, med Bryce Gervais som matchvinner. Storhamar vant hele 10-0 over Grüner, mens Frisk Asker og Sparta også tok seire i sine kamper.</p>
<p>Generated summary with Entity factor:</p> <p>Stavanger Oilers vant 2-1 over Vålerenga etter forlengning, med Bryce Gervais som matchvinner. Storhamar dominerte med en 10-0 seier over Grüner, der Tommy Hjelm og Patrick Thoresen begge scoret to mål. Frisk Asker snudde kampen og slo Lillehammer 4-3, mens Sparta knuste Ringerike 10-4 på bortebane.</p>

(c)

<p>Extracted Plots from news articles:</p> <p><=>Den amerikanske sentralbanken ser for seg to rentehopp i løpet av 2023, og øker anslaget på inflasjonen til 3,4 prosent i 2021.</s> <=>I mai alene hadde USA: kjerneinflasjon sin sterkeste årlige oppgang siden 1992, med en vekst på 3,8 prosent.</s> <=>Sentralbanken tror også den amerikanske økonomien vil vokse 7,0 prosent i år, opp fra deres tidligere estimat på 6,5 prosent.</s> <=>Estimater på arbeidsledigheten var det samme som før: 4,5 prosent.</e> <=>Den brede S&P 500-indeksen var for eksempel ned 0,28 prosent for dagen.</s></p>
--

(b)

- Plot plays a crucial role in personalized summarization, but it may introduce noise into pre-trained models, potentially diminishing the quality of generated summaries.
- User interested entities can be effectively captured by pre-trained models when generating personalized summaries in the topic-centric PersonalSum.

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.



Conclusion

- We propose a novel user-subjective guided [personalized summarization dataset](#), PersonalSum, with rich attributes.
 - This dataset features high-quality personalized summaries alongside their sources, user profiles, document-grounded question-answer pairs with answers' sources, and manually corrected machine-generated summaries with their corresponding sources.
- We highlight the [differences](#) between [LLM-generated summaries](#) and [human-annotated personalized summaries](#).
- We investigate and validate the impact of different [personalized signals](#) that may affect the performance of pretrained LLMs on personalized summarization task.



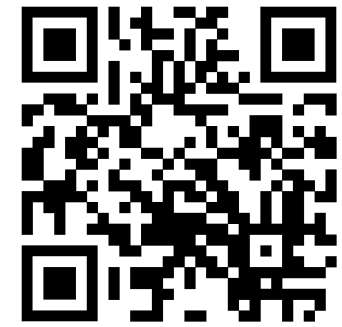
Some Findings

- As the number of user's historical annotations in the prompt increases, performance decreases. Possibly due to the scattered features of interest in users' history.
- Explicitly incorporating diverse personalized signals into the prompt affects the model's output to varying extents.
- Personalized signals play a crucial role in personalized summarization, but they may introduce noise into pre-trained models, potentially diminishing the quality of generated summaries.





Thank you!



Our paper: <https://arxiv.org/abs/2410.03905>

Github: <https://github.com/SmartmediaAI/PersonalSum/>

NorwAI

Norwegian Research Center
for AI Innovation

