

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



Norwegian Research Center for Al Innovation

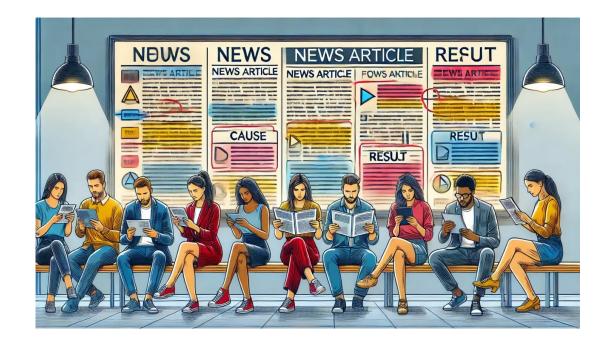




Centre for Centre for Research-based Innovation

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



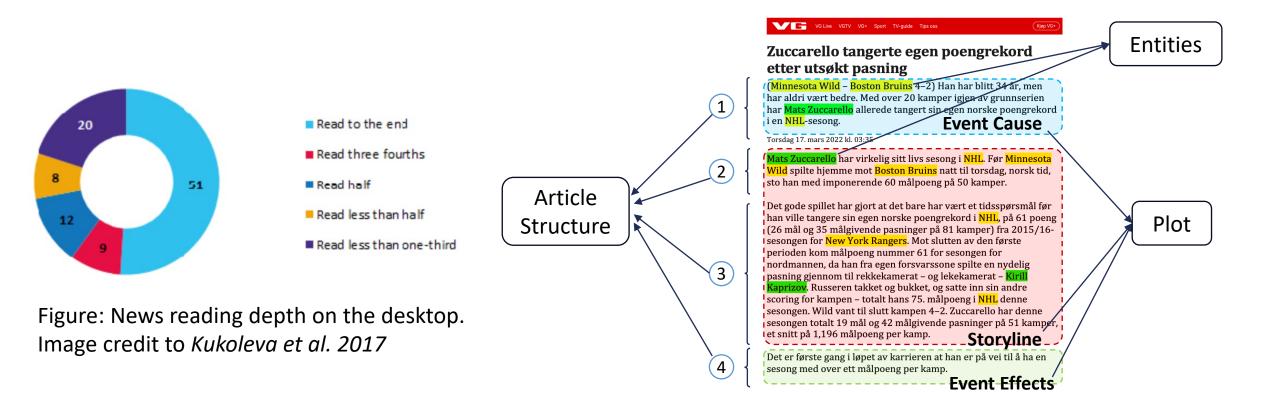
Kukoleva Olesya, Anna Preobrazhenskaya, and Olga Sidorova. 2017. *Media use habits: what, why, when, and how people read online*. UXMatters.





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?



NorwAl

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.

D		D .		Constr	ruction	User	Summary	D	
Datasets	Language Domain	Domain	#Summaries	Human Annotation	Multi- annotation	Profile	Source	Personalized	
		G	eneric Summariz	ation Datasets					
CNN/DM [2]	English	News	311,971	\checkmark	×	×	×	×	
XSum [3]	English	News	226,711	\checkmark	×	×	×	×	
NewsRoom [4]	English	News	1,212,740	\checkmark	×	×	×	×	
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	\checkmark	×	×	×	×	
WikiHow [8]	English	WikiHow	230,843	×	×	×	×	×	
		Con	trollable Summa	rization Datase	ts				
DUC [9]	English	News	300	\checkmark	\checkmark	×	×	×	
QMSum [10]	English	Meetings	1,808	\checkmark	\checkmark	×	×	×	
WikiAsp [11]	English	Wikipedia	566,881	×	\checkmark	×	×	×	
MACSUM [12]	English	News&Meetings	8333	\checkmark	\checkmark	×	\checkmark	×	
		Pers	onalized Summa	rization Datase	ts				
Amazon Reviews [13]	English	E-commerce	571,540,000	\checkmark	\checkmark	\checkmark	×	\checkmark	
PENS [14]	English	News Headline	20,600	\checkmark	\checkmark	×	×	\checkmark	
PersonalSum (ours)	Norwegian	News	1,816	\checkmark			\checkmark	\checkmark	

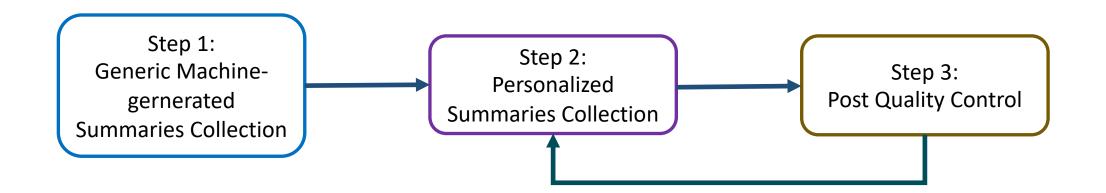
Table 1: Comparison between PersonalSum and existing popular summarization datasets.



PersonalSum: A User-Subjective Guided Personalized Summarization Dataset







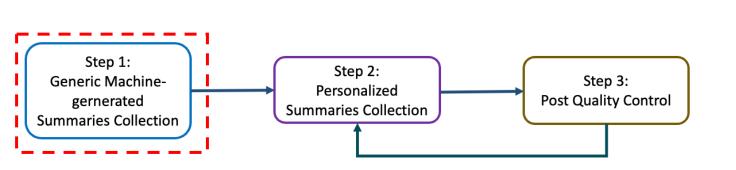




- Step 1: Collection of Machine-generated Generic Summaries
 - Some statistics: 465 news articles, 10 categories
 - Generated summaries: GPT-4

NorwAl

- 3 Norwegian students for quality control: 100% inter-agreement
- Annotated attributes: summaries, document-grounded question answers, sources



• • •								
<u> </u>	Title: Johannes Høsflot Klæbo kjøpte nabohuset for 13,3 millioner kroner Newsroom: ap							
•••	Creation_date: 2							
Dourslaad GPT-Questions Dourslaad GPT-Summary	Body: Johannes bekrefte kjøpet. J på Byksen i 2015 mest kjønt som la billedkunstner og Dette er høy pris Trondheim. – De visst oppussingal beholde fra da b beholgene på Bykå Hasilot sitt eget: utenlands. Kjøpe Konkurransesses heleid av VGTV gjør oppdrag for							
	Question	Hvor mye betalte Johannes Høsflot Klæbo for nabohuset sitt, og hva var opprinnelig prisen?	Generate Question					
Set API KEY	Answer	Johannes Klæbo betalte 13,3 millioner kroner, og den opprinnelige prisen ble satt til 11,2 millioner kroner.						
UI Scaling:	Source	Johannes Høsflot Klæbos nabohus lå ute til salgs for 11,2 millioner. Det ble solgt for 13,3 millioner.	ADD					
	Enter filename here.		Save to file					



- **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





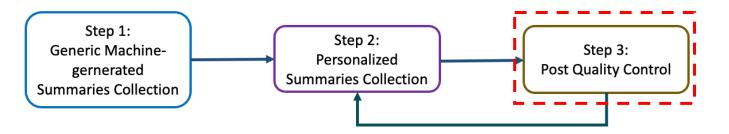


- Step 2: Collection of Personalized Summaries
 - Automatic filtering rules:
 - Summary length: >50 words
 - Task duration: >5 minutes

Veikommen til oppgaven! Vennligst les teksten nedenfor nøye. Din oppgave er å velge det korrekte spørsmål-svar paret som tilsvarer teksten og gi en oppsummering på to til tre setninger av den delen i teksten du syntes var mest inter Oppsumeringen skal ikke være på jeg form, og ikke inneholde annen tekst enn bare oppsummeringen. I kilder feltet skal du kopiere inn den delen av nyhetsartikkelen du oppsumerte, har du eksempelvis oppsumert fra de første tre setr nyhetsartikkelen skal du kopiere inn disse. Her er det viktig at du ikke skriver inn setninger som ikke finnes i den originale teksten. Teksten din vil bli evaluert basert på nøyaktigheten av svarene dine. For å unngå avvisning, er det viktig at du svarer korrekt på kontrollspørsmålene og velger et spørsmål-svar par som nøyaktig reflekterer informasjonen i teksten. Ukor svar kan føre til at bidraget ditt blir avvist. Du kan kvalifierer deg for en bonus basert på kvaliteten og kvantiteten av arbeidet ditt. Høykvalitetsbidrag som viser en grundig forståelse av teksten og et presist valg av spørsmål-svar par, vil øke sjansene dine for å motta en bonus . fullfører med høy kvalitet, desto større er sjansen for bonus. Tekst 1 (Minnesota Wild – Montréal Canadiens 8–2) Mats Zuccarelio (34) er i sitt livs form og noterte seg for to nye målpoeng natt til tirsdag. Nå har nordmannen målpoeng i ni kamper på rad. Helt på tampen av den andre perioden scoret Mati mål for sesongen. Nordmannen fikk ikke den beste pasningen fra rekkekamerat Kirill Kaprizov, men fikk lagt pucken til rette med foten. Fra nærmest død vinkel skjøt han smart opp i skulderen til keeper Cayden Primeau og i mål. Wild g en 5–1-ledelse, og njørmelagets storspill fortsatte i sluttperioden. Der noterte Zuccarelio og samme Kaprizov seg for en målgivende pasning da Matt Boldy la på ti 6–2, og kampen endte til slutt med en overbevisende & 2-æiert til Wild	
svar kan føre til at bidraget ditt blir avvist. Du kan kvalifisere deg for en bonus basert på kvaliteten og kvantiteten av arbeidet ditt. Høykvalitetsbidrag som viser en grundig forståelse av teksten og et presist valg av spørsmål-svar par, vil øke sjansene dine for å motta en bonus. fullfører med høy kvalitet, desto større er sjansen for bonus. Tekst 1 (Minnesota Wild – Montréal Canadiens 8–2) Mats Zuccarelio (34) er i sitt livs form og noterte seg for to nye målpoeng natt til tirsdag. Nå har nordmannen målpoeng i ni kamper på rad. Helt på tampen av den andre perioden scoret Mats mål for sesongen. Nordmannen fikk ikke den beste pasningen fra rekkekamerat Kirll Kaprizov, men fikk lagt pucken til rette med foten. Fra nærmest død vinkel skjøt han smart opp i skulderen til keeper Cayden Primeau og i mål. Wild g	
fullfører med høy kvalitet, desto større er sjansen for bonus. Tekst 1 (Minnesota Wild – Montréal Canadiens 8–2) Mats Zuccarello (34) er i sitt livs form og noterte seg for to nye målpoeng natt til tirsdag. Nå har nordmannen målpoeng i ni kamper på rad. Helt på tampen av den andre perioden scoret Mats mål for sesongen. Nordmannen fikk ikke den beste pasningen fra rekkekamerat Kirll Kaprizov, men fikk lagt pucken til rette med foten. Fra nærmest død vinkel skjøt han smart opp i skulderen til keeper Cayden Primeau og i mål. Wild g	Ukorrekte eller irrele
(Minnesota Wild – Montréal Canadiens 8–2) Mats Zuccarello (34) er i sitt livs form og noterte seg for to nye målpoeng natt til tirsdag. Nå har nordmannen målpoeng i ni kamper på rad. Helt på tampen av den andre perioden scoret Matt mål for sesongen. Nordmannen fikk ikke den beste pasningen fra rekkekamerat Kirll Kaprizov, men fikk lagt pucken til rette med foten. Fra nærmest død vinkel skjøt han smart opp i skulderen til keeper Cayden Primeau og i mål. Wild ç	us. Jo flere oppgave
mål for sesongen. Nordmannen fikk ikke den beste pasningen fra rekkekamerat Kirill Kaprizov, men fikk lagt pucken til rette med foten. Fra nærmest død vinkel skjat han smart opp i skulderen til keeper Cayden Primeau og i mål. Wid g	
målpoeng (12 scorlinger og 26 målgivende pasninger) på 32 kamper, et snitt på 1,19 poeng per kamp,34 åringen har aldri tildligere avsluttet en sesong med et snitt på over ett poeng per kamp,34 kamper passert di sesongene han hadde i Wild (37 poeng på 65 kamper i 2019/20 og 35 poeng på 42 kamper i 2020/21). Zuccarellos poengrekord i en grunnspillsesong er fra 2015/16, da han endte på 61 målpoeng for New York Rangers.— Jeg har god Vi er et virkelig godt lag, og når laget spiller bra, så synes jeg alle spiller bra, så zuccarello IS sar Tribune før tirsdagens kamp. Kun Kaprizov (17 mål og 30 målgivende pasninger) har flere målpoeng en Zuccarello to Minesola-laget Kaprizov hadde to målgivende pasninger i nattens kamp og nådde dermed 100 målpoeng i NHL i karrieren. Det har han klart på kun 92 kamper. I likhet med Zuccarello, har Kaprizov målpoeng i ni kamper på rad. Russeren har 17 målp kampene.Den neste kampen for Zuccarello og Wild er borte mot gamleklubben New York Rangers 28. januar. Wild ligger nå på fjerdeplass i Western Conference, men kan ved seier i lagets hengekamper ta seg opp på annenplass bak	gode spillere rundt m aget denne sesonger målpoeng på de siste
Velg korrekt spørsmål-svar par som tilhører teksten:	
Velg korrekt spørsmål-svar par som tilhører teksten: O 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 : 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. 	



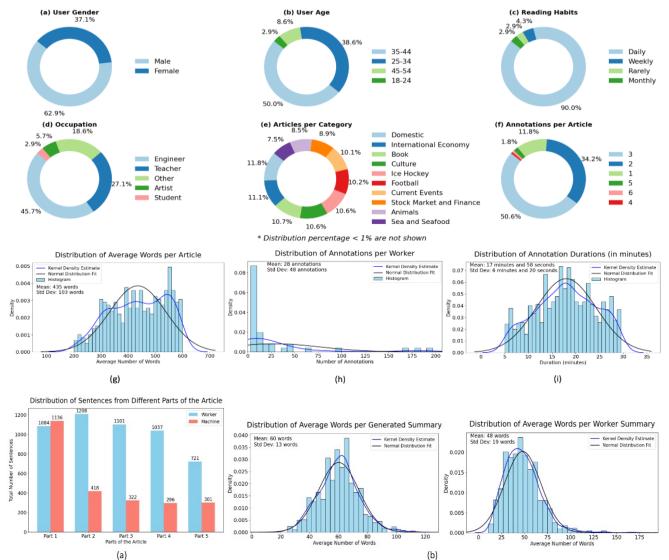
- Step 3: Post Quality Control
 - LLM evaluation:
 - Evaluate the annotated summary to a given article as well as its sources
 - Metrics: Coherence, Consistency, and Relevance
 - Human evaluation:
 - 10% of random sampled data for relevance score > 0.8
 - All annotations for relevance score <= 0.8







Data Statistics



- Machine-generated summaries are relatively longer than humanannotated 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, NorwAl-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-2 Rouge-L BERTScore	$37.90 \\ \pm 14.73$	$\begin{array}{c} 38.01 \\ \pm 14.82 \end{array}$	$\begin{array}{c} 37.56 \\ \pm 15.23 \end{array}$	$\begin{array}{c} 36.90 \\ \pm 16.25 \end{array}$	$\begin{array}{c} 37.93 \\ \pm 15.38 \end{array}$	$\begin{array}{c} 37.93 \\ \pm 15.36 \end{array}$	$38.03 \\ \pm 15.10$	$\frac{38.16}{\pm 15.43}$	38.43 ±15.22	
	Rouge-2	$\begin{array}{c} 17.00 \\ \pm 13.04 \end{array}$	$\begin{array}{c} 17.17 \\ \pm 13.19 \end{array}$	$\begin{array}{c} 16.89 \\ \pm 13.22 \end{array}$	$\begin{array}{c} 16.55 \\ \pm 13.52 \end{array}$	$\begin{array}{c} 17.05 \\ \pm 13.41 \end{array}$	17.06 ± 13.27	$\tfrac{17.27}{\pm 13.48}$	$17.20 \\ \pm 13.71$	17.47 ±13.65
	26.84 ± 13.10	$\begin{array}{c} 27.16 \\ \pm 13.13 \end{array}$	$\begin{array}{c} 26.85 \\ \pm 13.31 \end{array}$	$\begin{array}{c} 26.28 \\ \pm 14.05 \end{array}$	27.15 ± 13.65	26.96 ± 13.49	$\frac{27.37}{\pm 13.53}$	27.37 ± 13.74	27.45 ±13.71	
	BERTScore	$\begin{array}{c} 75.00 \\ \pm 5.39 \end{array}$	75.16 ± 5.30	$\begin{array}{c} 74.76 \\ \pm 5.72 \end{array}$	$\begin{array}{c} 74.64 \\ \pm 6.14 \end{array}$	$\begin{array}{c} 74.98 \\ \pm 5.79 \end{array}$	$\begin{array}{c} 75.00 \\ \pm 5.64 \end{array}$	$\begin{array}{c} 75.02 \\ \pm 5.62 \end{array}$	$\frac{75.20}{\pm 5.64}$	75.20 ±5.60
Gemini 1.0 Pro	Rouge-1	$35.21 \\ \pm 13.51$	$\tfrac{35.67}{\pm 14.09}$	$\begin{array}{c} 35.30 \\ \pm 13.46 \end{array}$	$\substack{35.45\\\pm13.45}$	$\begin{array}{c} 35.42 \\ \pm 13.97 \end{array}$	$35.60 \\ \pm 13.89$	35.91 ±14.03	$35.62 \\ \pm 14.05$	$35.47 \\ \pm 13.87$
	Rouge-2	$14.32 \\ \pm 11.14$	$\tfrac{14.76}{\pm 11.60}$	$\begin{array}{c} 14.27 \\ \pm 11.09 \end{array}$	$\begin{array}{c} 14.37 \\ \pm 11.02 \end{array}$	14.55 ± 11.51	$14.42 \\ \pm 11.34$	14.88 ±11.70	$\begin{array}{c} 14.70 \\ \pm 11.62 \end{array}$	$\begin{array}{c} 14.59 \\ \pm 11.43 \end{array}$
	Rouge-L	$25.21 \\ \pm 11.86$	$\tfrac{25.75}{\pm 12.52}$	$\begin{array}{c} 25.18 \\ \pm 11.82 \end{array}$	$\begin{array}{c} 25.57 \\ \pm 11.91 \end{array}$	$\begin{array}{c} 25.46 \\ \pm 12.22 \end{array}$	25.55 ± 12.17	25.86 ±12.38	25.53 ± 12.33	25.27 ± 11.90
	BERTScore	$\begin{array}{c} 74.52 \\ \pm 5.07 \end{array}$	74.74 ±5.24	74.42 ± 5.02	74.56 ± 5.01	$74.49 \\ \pm 5.28$	$\tfrac{74.63}{\pm 5.14}$	$74.53 \\ \pm 5.36$	$74.54 \\ \pm 5.28$	$\begin{array}{c} 74.41 \\ \pm 5.28 \end{array}$
	Rouge-1	$\begin{array}{c} 33.88 \\ \pm 12.62 \end{array}$	$\begin{array}{c} 34.14 \\ \pm 13.54 \end{array}$	$\begin{array}{c} 34.01 \\ \pm 13.58 \end{array}$	$\begin{array}{c} 33.83 \\ \pm 13.42 \end{array}$	$\begin{array}{c} 33.96 \\ \pm 13.31 \end{array}$	34.15 ± 13.60	$33.81 \\ \pm 13.49$	$\frac{34.24}{\pm 13.88}$	34.29 ±13.66
NorwAI- Mixtral-	Rouge-2	$13.36 \\ \pm 10.43$	$\begin{array}{c} 13.66 \\ \pm 11.13 \end{array}$	$\tfrac{13.77}{\pm 11.00}$	$\begin{array}{c} 13.56 \\ \pm 11.13 \end{array}$	$\begin{array}{c} 13.69 \\ \pm 10.95 \end{array}$	13.75 ± 11.05	$13.62 \\ \pm 10.96$	13.77 ± 11.26	13.89 ±11.03
8x7B- instruct	Rouge-L	$23.58 \\ \pm 10.49$	$\begin{array}{c} 24.12 \\ \pm 11.49 \end{array}$	$\begin{array}{c} 24.03 \\ \pm 11.38 \end{array}$	$\begin{array}{c} 23.99 \\ \pm 11.55 \end{array}$	$\begin{array}{c} 24.01 \\ \pm 11.18 \end{array}$	24.16 ±11.51	24.04 ± 11.26	24.04 ± 11.73	$\tfrac{\underline{24.13}}{\pm 11.28}$
	BERTScore	$73.51 \\ \pm 4.72$	$\begin{array}{c} 73.69 \\ \pm 4.95 \end{array}$	$\begin{array}{c} 73.79 \\ \pm 4.94 \end{array}$	$\begin{array}{c} 73.79 \\ \pm 4.92 \end{array}$	$\tfrac{73.84}{\pm 4.82}$	$\begin{array}{c} 73.78 \\ \pm 5.09 \end{array}$	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

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	A11
GPT-3.5 Turbo Rouge-2 Rouge-L BERTScore	Rouge-1	$37.90 \\ \pm 14.73$	$\begin{array}{c} 38.01 \\ \pm 14.82 \end{array}$	$\begin{array}{c} 37.56 \\ \pm 15.23 \end{array}$	$\begin{array}{c} 36.90 \\ \pm 16.25 \end{array}$	$\begin{array}{c} 37.93 \\ \pm 15.38 \end{array}$	$\begin{array}{c} 37.93 \\ \pm 15.36 \end{array}$	$38.03 \\ \pm 15.10$	$\frac{38.16}{\pm 15.43}$	38.43 ±15.22
	Rouge-2	$\begin{array}{c} 17.00 \\ \pm 13.04 \end{array}$	$\begin{array}{c} 17.17 \\ \pm 13.19 \end{array}$	$\begin{array}{c} 16.89 \\ \pm 13.22 \end{array}$	$\begin{array}{c} 16.55 \\ \pm 13.52 \end{array}$	$\begin{array}{c} 17.05 \\ \pm 13.41 \end{array}$	17.06 ± 13.27	$\tfrac{17.27}{\pm 13.48}$	17.20 ± 13.71	17.47 ±13.65
	$\begin{array}{c} 26.84 \\ \pm 13.10 \end{array}$	$\begin{array}{c} 27.16 \\ \pm 13.13 \end{array}$	$\begin{array}{c} 26.85 \\ \pm 13.31 \end{array}$	$\begin{array}{c} 26.28 \\ \pm 14.05 \end{array}$	$\begin{array}{c} 27.15 \\ \pm 13.65 \end{array}$	26.96 ± 13.49	$\frac{27.37}{\pm 13.53}$	27.37 ± 13.74	27.45 ±13.71	
	BERTScore	$\begin{array}{c} 75.00 \\ \pm 5.39 \end{array}$	$\begin{array}{c} 75.16 \\ \pm 5.30 \end{array}$	$\begin{array}{c} 74.76 \\ \pm 5.72 \end{array}$	$\begin{array}{c} 74.64 \\ \pm 6.14 \end{array}$	74.98 ± 5.79	$\begin{array}{c} 75.00 \\ \pm 5.64 \end{array}$	$\begin{array}{c} 75.02 \\ \pm 5.62 \end{array}$	$\frac{75.20}{\pm 5.64}$	75.20 ±5.60
Gemini 1.0 Pro	Rouge-1	$\begin{array}{c} 35.21 \\ \pm 13.51 \end{array}$	$\tfrac{35.67}{\pm 14.09}$	$\begin{array}{c} 35.30 \\ \pm 13.46 \end{array}$	$\substack{35.45\\\pm13.45}$	$\begin{array}{c} 35.42 \\ \pm 13.97 \end{array}$	$35.60 \\ \pm 13.89$	35.91 ±14.03	$35.62 \\ \pm 14.05$	$35.47 \\ \pm 13.87$
	Rouge-2	$\begin{array}{c} 14.32 \\ \pm 11.14 \end{array}$	$\tfrac{14.76}{\pm 11.60}$	$\begin{array}{c} 14.27 \\ \pm 11.09 \end{array}$	$\begin{array}{c} 14.37 \\ \pm 11.02 \end{array}$	$\begin{array}{c} 14.55 \\ \pm 11.51 \end{array}$	$\begin{array}{c} 14.42 \\ \pm 11.34 \end{array}$	14.88 ±11.70	$\begin{array}{c} 14.70 \\ \pm 11.62 \end{array}$	$14.59 \\ \pm 11.43$
	Rouge-L	$\begin{array}{c} 25.21 \\ \pm 11.86 \end{array}$	$\tfrac{25.75}{\pm 12.52}$	$\begin{array}{c} 25.18 \\ \pm 11.82 \end{array}$	$\begin{array}{c} 25.57 \\ \pm 11.91 \end{array}$	$\begin{array}{c} 25.46 \\ \pm 12.22 \end{array}$	25.55 ± 12.17	25.86 ±12.38	25.53 ± 12.33	25.27 ± 11.90
	BERTScore	$\begin{array}{c} 74.52 \\ \pm 5.07 \end{array}$	74.74 ±5.24	$\begin{array}{c} 74.42 \\ \pm 5.02 \end{array}$	74.56 ± 5.01	74.49 ± 5.28	$\tfrac{74.63}{\pm 5.14}$	$74.53 \\ \pm 5.36$	74.54 ± 5.28	74.41 ± 5.28
	Rouge-1	$\begin{array}{c} 33.88 \\ \pm 12.62 \end{array}$	$\begin{array}{c} 34.14 \\ \pm 13.54 \end{array}$	$\begin{array}{c} 34.01 \\ \pm 13.58 \end{array}$	$\begin{array}{c} 33.83 \\ \pm 13.42 \end{array}$	$\begin{array}{c} 33.96 \\ \pm 13.31 \end{array}$	$\begin{array}{c} 34.15 \\ \pm 13.60 \end{array}$	$33.81 \\ \pm 13.49$	$\frac{34.24}{\pm 13.88}$	34.29 ±13.66
NorwAI- Mixtral-	Rouge-2	$\begin{array}{c} 13.36 \\ \pm 10.43 \end{array}$	$\begin{array}{c} 13.66 \\ \pm 11.13 \end{array}$	$\tfrac{13.77}{\pm 11.00}$	$\begin{array}{c} 13.56 \\ \pm 11.13 \end{array}$	$\begin{array}{c} 13.69 \\ \pm 10.95 \end{array}$	13.75 ± 11.05	$\begin{array}{c} 13.62 \\ \pm 10.96 \end{array}$	13.77 ± 11.26	13.89 ±11.03
8x7B- instruct	Rouge-L	$\begin{array}{c} 23.58 \\ \pm 10.49 \end{array}$	$\begin{array}{c} 24.12 \\ \pm 11.49 \end{array}$	$\begin{array}{c} 24.03 \\ \pm 11.38 \end{array}$	$\begin{array}{c} 23.99 \\ \pm 11.55 \end{array}$	$\begin{array}{c} 24.01 \\ \pm 11.18 \end{array}$	24.16 ±11.51	$\begin{array}{c} 24.04 \\ \pm 11.26 \end{array}$	24.04 ± 11.73	$\tfrac{\underline{24.13}}{\pm 11.28}$
	BERTScore	$\begin{array}{c} 73.51 \\ \pm 4.72 \end{array}$	$\begin{array}{c} 73.69 \\ \pm 4.95 \end{array}$	$\begin{array}{c} 73.79 \\ \pm 4.94 \end{array}$	$\begin{array}{c} 73.79 \\ \pm 4.92 \end{array}$	$\tfrac{73.84}{\pm 4.82}$	$\begin{array}{c} 73.78 \\ \pm 5.09 \end{array}$	$73.75 \\ \pm 4.94$	73.77 ± 4.98	73.95 ±4.87

Few overlapped NEs in one HIT!

- 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
<u> </u>	Rouge-1	$\begin{array}{c} 37.14 \\ \pm 14.02 \end{array}$	$\begin{array}{c} 39.16 \\ \pm 14.13 \end{array}$	$\begin{array}{c} 39.21 \\ \pm 14.36 \end{array}$	$\frac{39.42}{\pm 15.09}$	$\begin{array}{c} 39.37 \\ \pm 14.18 \end{array}$	38.78 ± 14.36	39.90 ±14.42	$39.30 \\ \pm 14.01$	$\begin{array}{c} 39.07 \\ \pm 14.46 \end{array}$
	Rouge-2	$\begin{array}{c} 16.43 \\ \pm 12.06 \end{array}$	$\begin{array}{c} 17.58 \\ \pm 12.30 \end{array}$	$\begin{array}{c} 17.94 \\ \pm 12.78 \end{array}$	$\begin{array}{c} 18.15 \\ \pm 12.84 \end{array}$	$\begin{array}{c} 17.94 \\ \pm 12.42 \end{array}$	$17.86 \\ \pm 12.43$	18.81 ±12.75	$\frac{18.22}{\pm 12.33}$	$\begin{array}{c} 17.61 \\ \pm 12.35 \end{array}$
Turbo	Rouge-L	$26.43 \\ \pm 12.30$	27.23 ± 12.20	27.72 ± 12.86	$28.03 \\ \pm 13.26$	27.70 ± 12.40	27.27 ± 12.37	$\frac{28.06}{\pm 12.56}$	28.12 ±12.67	$27.50 \\ \pm 12.54$
	BERTScore	$\begin{array}{c} 74.84 \\ \pm 5.13 \end{array}$	75.57 ± 5.20	$75.60 \\ \pm 5.38$	$75.59 \\ \pm 5.59$	$75.66 \\ \pm 5.25$	$75.51 \\ \pm 5.30$	75.75 ±5.37	$\frac{75.73}{\pm 5.13}$	$\begin{array}{c} 75.54 \\ \pm 5.32 \end{array}$
	Rouge-1	$\begin{array}{c} 35.46 \\ \pm 12.84 \end{array}$	36.48 ±12.96	$36.30 \\ \pm 13.25$	$35.26 \\ \pm 14.32$	$\begin{array}{c} 35.14 \\ \pm 14.53 \end{array}$	$\frac{36.46}{\pm 14.10}$	$36.32 \\ \pm 13.54$	$36.18 \\ \pm 13.86$	$\begin{array}{c} 36.33 \\ \pm 13.81 \end{array}$
Gemini	Rouge-2	$\begin{array}{c} 14.52 \\ \pm 11.02 \end{array}$	$\begin{array}{c} 14.74 \\ \pm 10.41 \end{array}$	$\begin{array}{c} 15.04 \\ \pm 10.35 \end{array}$	$\begin{array}{c} 13.89 \\ \pm 10.46 \end{array}$	14.24 ± 10.93	15.19 ±11.26	$14.99 \\ \pm 11.38$	$\frac{15.11}{\pm 11.22}$	15.06 ± 11.11
1.0 Pto	1.0 Pro Rouge-L	$\begin{array}{c} 24.88 \\ \pm 11.16 \end{array}$	$\begin{array}{c} 25.74 \\ \pm 11.59 \end{array}$	25.73 ± 11.17	$\begin{array}{c} 24.54 \\ \pm 12.20 \end{array}$	$24.63 \\ \pm 11.95$	25.88 ±12.56	25.75 ± 12.17	25.59 ± 12.42	$\frac{25.76}{\pm 11.89}$
В	BERTScore	$74.45 \\ \pm 4.95$	$\tfrac{74.82}{\pm 4.98}$	$\begin{array}{c} 74.66 \\ \pm 5.20 \end{array}$	$\begin{array}{c} 74.32 \\ \pm 5.61 \end{array}$	$\begin{array}{c} 73.85 \\ \pm 6.41 \end{array}$	74.85 ±5.49	$74.72 \\ \pm 5.44$	$\begin{array}{c} 74.60 \\ \pm 5.54 \end{array}$	$\begin{array}{c} 74.76 \\ \pm 5.11 \end{array}$
	Rouge-1	$\begin{array}{c} 32.80 \\ \pm 11.98 \end{array}$	$\tfrac{33.87}{\pm 12.14}$	$\begin{array}{c} 33.20 \\ \pm 11.85 \end{array}$	$\begin{array}{c} 33.23 \\ \pm 13.19 \end{array}$	$\begin{array}{c} 33.65 \\ \pm 12.49 \end{array}$	$\frac{55.19}{\pm 12.57}$	34.32 ±12.90	$33.58 \\ \pm 12.71$	$\begin{array}{c} 33.34 \\ \pm 13.17 \end{array}$
NorwAI- Mixtral-	Rouge-2	$12.29 \\ \pm 9.47$	$\begin{array}{c} 13.28 \\ \pm 9.55 \end{array}$	$\begin{array}{c} 12.61 \\ \pm 8.87 \end{array}$	$\begin{array}{c} 13.07 \\ \pm 9.92 \end{array}$	$\begin{array}{c} 13.04 \\ \pm 9.65 \end{array}$	$12.92 \\ \pm 9.68$	13.82 ±10.22	$\begin{array}{c} 12.66 \\ \pm 9.31 \end{array}$	$\frac{13.35}{\pm 10.21}$
8x7B- instruct	Rouge-L	$\begin{array}{c} 22.66 \\ \pm 10.26 \end{array}$	$\tfrac{\underline{23.62}}{\pm 10.26}$	$\begin{array}{c} 22.68 \\ \pm 9.24 \end{array}$	$\begin{array}{c} 23.01 \\ \pm 10.78 \end{array}$	$\begin{array}{c} 23.45 \\ \pm 10.17 \end{array}$	22.96 ± 10.18	23.93 ±10.97	23.23 ± 10.14	$\begin{array}{c} 23.32 \\ \pm 10.82 \end{array}$
	BERTScore	73.15 ± 4.57	$\tfrac{73.77}{\pm 4.52}$	$\begin{array}{c} 73.63 \\ \pm 4.41 \end{array}$	73.73 ± 4.75	$\begin{array}{c} 73.60 \\ \pm 4.51 \end{array}$	$73.69 \\ \pm 4.72$	74.13 ±4.80	$73.60 \\ \pm 4.72$	$\begin{array}{c} 73.72 \\ \pm 4.80 \end{array}$
	Rouge-1	$\begin{array}{c} 35.98 \\ \pm 13.02 \end{array}$	$\begin{array}{c} 17.86 \\ \pm 14.97 \end{array}$	$\begin{array}{c} 16.73 \\ \pm 13.82 \end{array}$	17.88 ± 16.50	$18.30 \\ \pm 16.23$	17.86 ± 15.96	16.60 ± 14.99	$\begin{array}{c} 18.61 \\ \pm 16.45 \end{array}$	$\begin{array}{c} 18.05 \\ \pm 15.80 \end{array}$
Meta- Llama3-	Rouge-2	$\begin{array}{c} 14.15 \\ \pm 10.29 \end{array}$	$7.20 \\ \pm 8.88$	$\begin{array}{c} 6.74 \\ \pm 8.00 \end{array}$	7.33 ± 9.36	7.54 ± 9.61	$\begin{array}{c} 7.43 \\ \pm 8.98 \end{array}$	$\begin{array}{c} 6.48 \\ \pm 8.24 \end{array}$	$7.66 \\ \pm 9.41$	$7.29 \\ \pm 9.30$
70B- Instruct	Rouge-L	$\begin{array}{c} 24.25 \\ \pm 10.21 \end{array}$	$\begin{array}{c} 13.56 \\ \pm 10.86 \end{array}$	$\begin{array}{c} 12.87 \\ \pm 9.97 \end{array}$	$\begin{array}{c} 13.17 \\ \pm 11.67 \end{array}$	$\begin{array}{c} 13.49 \\ \pm 11.60 \end{array}$	$13.43 \\ \pm 11.35$	12.52 ± 10.95	$13.69 \\ \pm 11.63$	$\begin{array}{c} 13.33 \\ \pm 11.28 \end{array}$
	BERTScore	$\begin{array}{c} 74.10 \\ \pm 4.99 \end{array}$	$\begin{array}{c} 71.03 \\ \pm 4.62 \end{array}$	$\begin{array}{c} 69.93 \\ \pm 5.09 \end{array}$	$\begin{array}{c} 70.05 \\ \pm 6.50 \end{array}$	$\begin{array}{c} 70.48 \\ \pm 5.75 \end{array}$	$\begin{array}{c} 70.31 \\ \pm 6.11 \end{array}$	69.78 ± 6.10	$\begin{array}{c} 70.59 \\ \pm 5.64 \end{array}$	$\begin{array}{c} 70.03 \\ \pm 6.10 \end{array}$

- 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 NorwAl-Mixtral-8x7B-instruct
- Evaluation metrics: Consistency and Coherence
- Internal agreement: Fleiss' kappa (κ)

Models	Cons	istency / Fleiss' l	tappa	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	

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





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 NorwAl-Mixtral-8x7B-instruct: *Focus on Different Topics, Divergent Plot Emphasis, Incomplete Output*

Models	Cons	sistency / Fleiss' k	appa	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	

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



Case Study

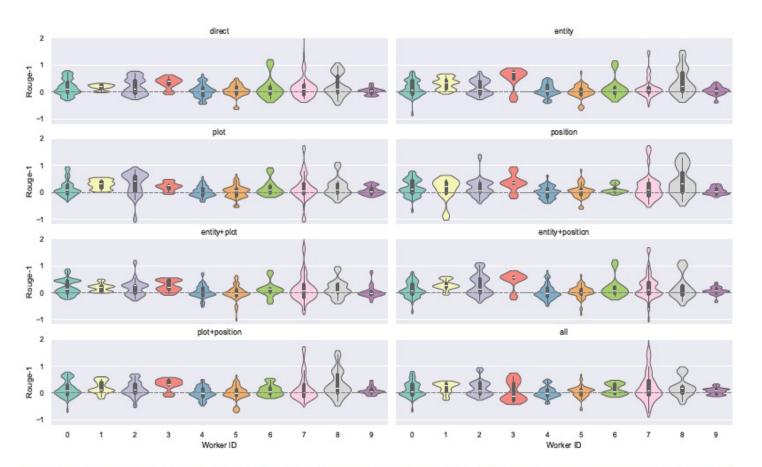


Figure 3: Experimental results showing improvements in the ROUGE-1 score from personalized prompting compared to generic summaries using GPT-3.5 Turbo for each worker. The X-axis represents worker IDs, and the Y-axis represents the ROUGE-1 score improvements.

• Two instances for further analysis w.r.t. ROUGE1 score on 5-shot prompting:

- Worker 1: higher improvements
- Worker 3: reduced improvements

NorwAl



Case Study

Previous extracted plots in worker's 5-shot summaries: Event Storyline (<s>), Event Result (<e>)

Worker summary:

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 opggangen 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 flaskehalser på tilbudssiden som bakgrunn for oppgangen..

Generated Generic summary:

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.

Generated summary with Plot factor:

Federal Reserve, den amerikanske sentralbanken, planlegger to renteøkninger 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,(<s>) 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.(<s>)

(a)

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, Sommartider, Stavanger Oilers, Evan Buitenhuis, Oilers, Trym Gran, Lillehammer, Jeppe Meyer, Andreas

Worker summary:

Stavanger Ollers vant 2-1 mot Välerenga 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 Välerenga før forlengningen. Samtidig vant Storhamar hele 10-0 over Grüner i Oslo.

Generated Generic summary:

Stavanger Ollers 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.

Generated summary with Entity factor:

Stavanger Ollers 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.

(c)

Extracted Plots from news articles:

<s>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 USAs kjerneinflasjon sin sterkeste årlige oppgang siden 1992, med en vekst på 3,8 prosent.</s>

<s>Sentralbanken tror også den amerikanske økonomien vil vokse 7,0 prosent i år, opp fra deres tidligere estimat på 6,5 prosent.</s> <e>Estimatet på arbeidsledigheten var det samme som før: 4,5 prosent.</e> <s>Den brede S&P 500-indeksen var for eksempel ned 0,28 prosent for dagen.</s>

(b)

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-40. 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.

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 topiccentric PersonalSum.



NorwAl

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, documentgrounded 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/SmartmediaAl/PersonalSum/

NorwAl

Norwegian Research Center for AI Innovation

