Supplementary Material: 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`

Contents

A Appendix 1
- A.1 Accessibility . . . 1
- A.2 Prompts . . . 1
- A.3 Supplementary results for PersonalSum . . . 4
- A.4 Supplementary results for Topic-centric PersonalSum . . . 5
- A.5 Error Case Analysis on PersonalSum . . . 5
- A.6 Statistics for Topic-centric PersonalSum . . . 5

B Datasheets for PersonalSum dataset 9
- B.1 Motivation . . . 9
- B.2 Composition . . . 10
- B.3 Collection Process . . . 11
- B.4 Preprocessing / cleaning / labeling . . . 14
- B.5 Uses . . . 14
- B.6 Distribution . . . 14
- B.7 Maintenance . . . 15

C Responsibility Statement 16

---

[*] Corresponding author

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

A Appendix

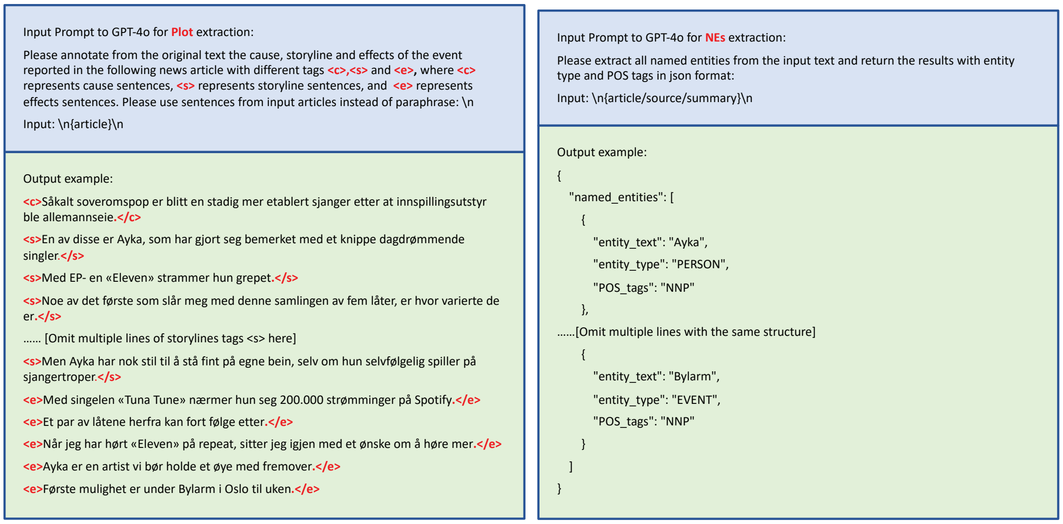
The appendix contains instructions on how to access the dataset, all prompt templates used in our paper, supplementary results for PersonalSum and Topic-centric PersonalSum, an error case analysis of PersonalSum, and the statistics for Topic-centric PersonalSum.

A.1 Accessibility

All the resources in our work are accessible online [2], including dataset card and related code for data collection and quality control on Amazon Turk. The licensing for the dataset is under a CC BY-NC 4.0[3]. We will consistently maintain and update the resources to ensure long-term usability.

A.2 Prompts

This section lists all prompts used in the paper. Figure 1(a) and (b) show the prompts used to extract plot and named entities (NEs) from the input article using GPT-4o, respectively. Figure 2 and Figure 3 respectively show the Norwegian prompts we adopted to generate generic summaries (Generic), personalized summaries with only user history (Direct), personalized summaries with user history and entity (Entity)/plot (Plot)/position (Position) factor for all models in our experiments. When testing different combinations of factors, we add the associated factors to the user's history. For example, when exploring all combinations of factors (All), the prompt is as shown in Figure 4. Figure 5 shows the prompt used to control the quality of annotated summary in Section 3.1.



Input Prompt to GPT-4o for **Plot** extraction:

Please annotate from the original text the cause, storyline and effects of the event reported in the following news article with different tags <c>,<s> and <e>, where <c> represents cause sentences, <s> represents storyline sentences, and <e> represents effects sentences. Please use sentences from input articles instead of paraphrase: \n

Input: \n{article}\n

Output example:

<c>Såkalt soveromspop er blitt en stadig mer etablert sjanger etter at innspillingsutstyr ble allemannseie.</c>

<s>En av disse er Ayka, som har gjort seg bemerket med et knippe dagdrømmende singler.</s>

<s>Med EP- en «Eleven» strammer hun grepet.</s>

<s>Noe av det første som slår meg med denne samlingen av fem låter, er hvor varierte de er.</s>

…… [Omit multiple lines of storylines tags <s> here]

<s>Men Ayka har nok stil til å stå fint på egne bein, selv om hun selvfølgelig spiller på sjangertroper.</s>

<e>Med singelen «Tuna Tune» nærmer hun seg 200.000 strømminger på Spotify.</e>

<e>Et par av låtene herfra kan fort følge etter.</e>

<e>Når jeg har hørt «Eleven» på repeat, sitter jeg igjen med et ønske om å høre mer.</e>

<e>Ayka er en artist vi bør holde et øye med fremover.</e>

<e>Første mulighet er under Bylarm i Oslo til uken.</e>

(a)

Input Prompt to GPT-4o for **NEs** extraction:

Please extract all named entities from the input text and return the results with entity type and POS tags in json format:

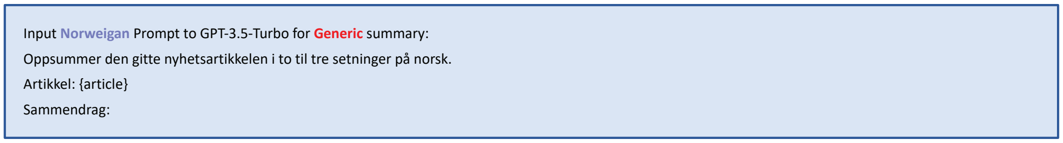
Input: \n{article/source/summary}\n

Output example:

```
{
   "named_entities": [
      {
         "entity_text": "Ayka",
         "entity_type": "PERSON",
         "POS_tags": "NNP"
      },
……[Omit multiple lines with the same structure]
      {
         "entity_text": "Bylarm",
         "entity_type": "EVENT",
         "POS_tags": "NNP"
      }
   ]
}
```

(b)

Figure 1: Prompt template using GPT-4o on (a) Plot extraction, and (b) Named Entity (NE) recognition.



Input **Norweigan** Prompt to GPT-3.5-Turbo for **Generic** summary:

Oppsummer den gitte nyhetsartikkelen i to til tre setninger på norsk.

Artikkel: {article}

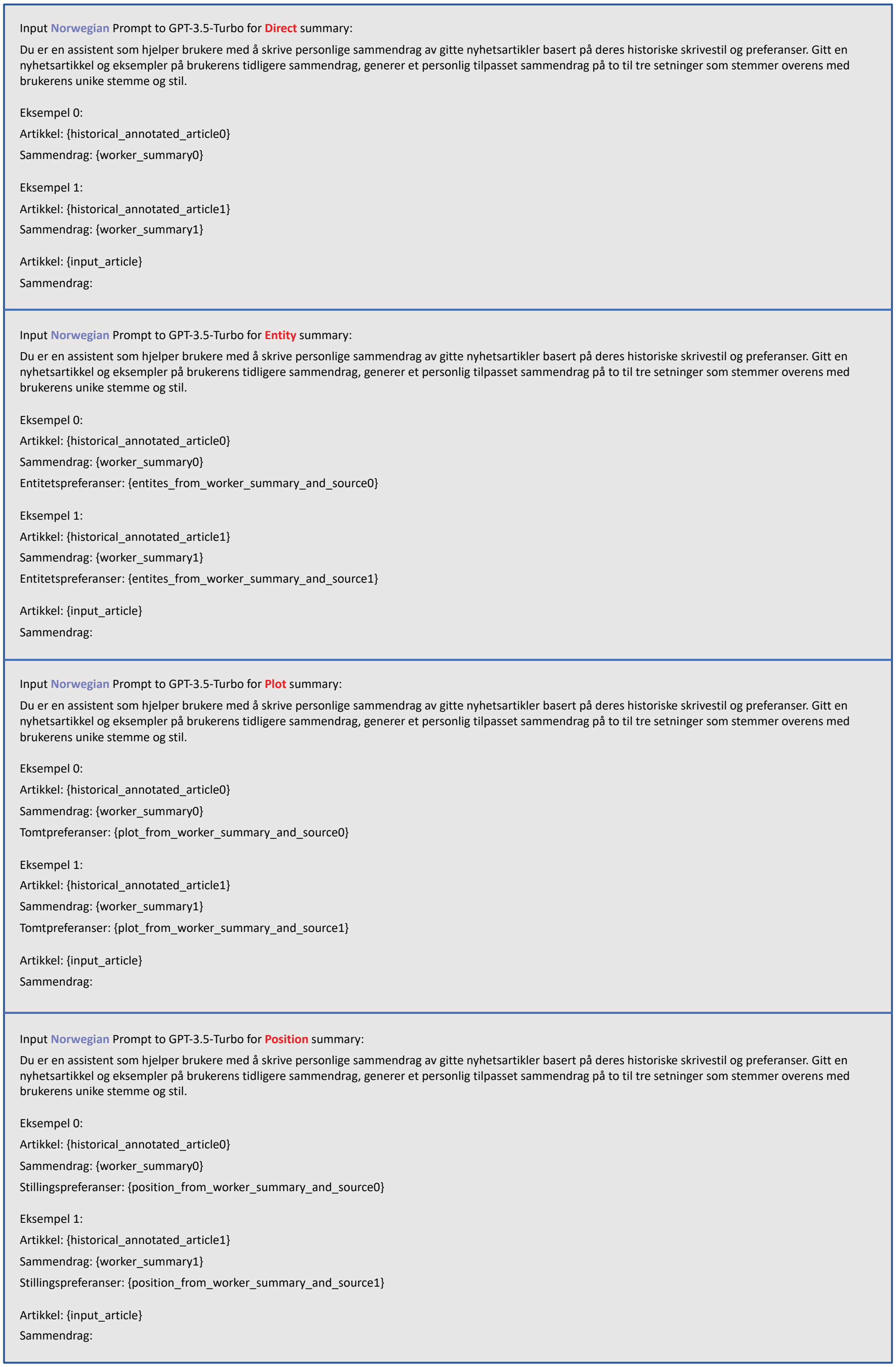
Sammendrag:

Figure 2: Prompt for generic summary generation.

---

[2] `https://github.com/SmartmediaAI/PersonalSum`, and `https://huggingface.co/datasets/PersonalLab/PersonalSum` with a Croissant metadata record.

[3] `https://creativecommons.org/licenses/by-nc/4.0/`



Input **Norwegian** Prompt to GPT-3.5-Turbo for **Direct** summary:

Du er en assistent som hjelper brukere med å skrive personlige sammendrag av gitte nyhetsartikler basert på deres historiske skrivestil og preferanser. Gitt en nyhetsartikkel og eksempler på brukerens tidligere sammendrag, generer et personlig tilpasset sammendrag på to til tre setninger som stemmer overens med brukerens unike stemme og stil.

Eksempel 0:
Artikkel: {historical_annotated_article0}
Sammendrag: {worker_summary0}

Eksempel 1:
Artikkel: {historical_annotated_article1}
Sammendrag: {worker_summary1}

Artikkel: {input_article}
Sammendrag:

Input **Norwegian** Prompt to GPT-3.5-Turbo for **Entity** summary:

Du er en assistent som hjelper brukere med å skrive personlige sammendrag av gitte nyhetsartikler basert på deres historiske skrivestil og preferanser. Gitt en nyhetsartikkel og eksempler på brukerens tidligere sammendrag, generer et personlig tilpasset sammendrag på to til tre setninger som stemmer overens med brukerens unike stemme og stil.

Eksempel 0:
Artikkel: {historical_annotated_article0}
Sammendrag: {worker_summary0}
Entitetspreferanser: {entites_from_worker_summary_and_source0}

Eksempel 1:
Artikkel: {historical_annotated_article1}
Sammendrag: {worker_summary1}
Entitetspreferanser: {entites_from_worker_summary_and_source1}

Artikkel: {input_article}
Sammendrag:

Input **Norwegian** Prompt to GPT-3.5-Turbo for **Plot** summary:

Du er en assistent som hjelper brukere med å skrive personlige sammendrag av gitte nyhetsartikler basert på deres historiske skrivestil og preferanser. Gitt en nyhetsartikkel og eksempler på brukerens tidligere sammendrag, generer et personlig tilpasset sammendrag på to til tre setninger som stemmer overens med brukerens unike stemme og stil.

Eksempel 0:
Artikkel: {historical_annotated_article0}
Sammendrag: {worker_summary0}
Tomtpreferanser: {plot_from_worker_summary_and_source0}

Eksempel 1:
Artikkel: {historical_annotated_article1}
Sammendrag: {worker_summary1}
Tomtpreferanser: {plot_from_worker_summary_and_source1}

Artikkel: {input_article}
Sammendrag:

Input **Norwegian** Prompt to GPT-3.5-Turbo for **Position** summary:

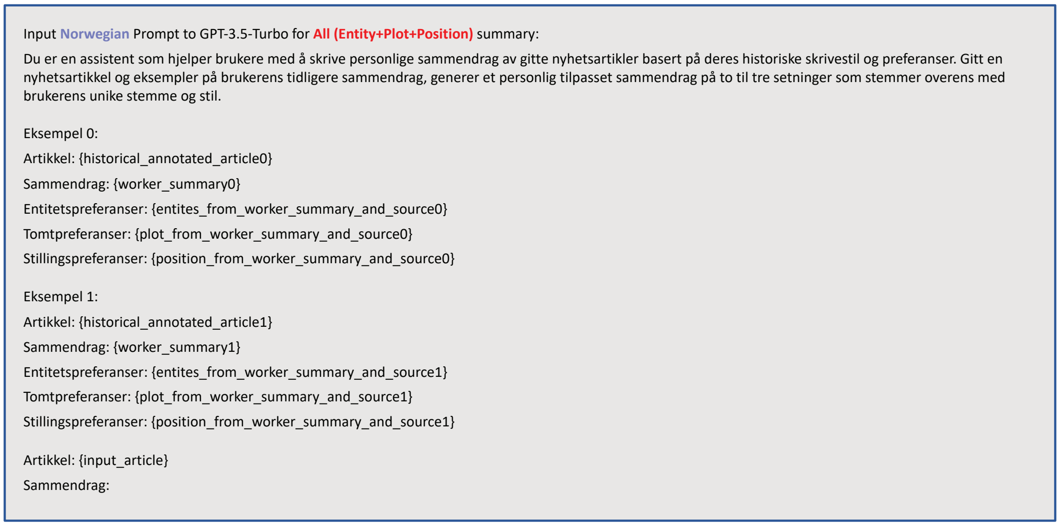
Du er en assistent som hjelper brukere med å skrive personlige sammendrag av gitte nyhetsartikler basert på deres historiske skrivestil og preferanser. Gitt en nyhetsartikkel og eksempler på brukerens tidligere sammendrag, generer et personlig tilpasset sammendrag på to til tre setninger som stemmer overens med brukerens unike stemme og stil.

Eksempel 0:
Artikkel: {historical_annotated_article0}
Sammendrag: {worker_summary0}
Stillingspreferanser: {position_from_worker_summary_and_source0}

Eksempel 1:
Artikkel: {historical_annotated_article1}
Sammendrag: {worker_summary1}
Stillingspreferanser: {position_from_worker_summary_and_source1}

Artikkel: {input_article}
Sammendrag:

Figure 3: Prompts for generating personalized summaries.



Input **Norwegian** Prompt to GPT-3.5-Turbo for **All (Entity+Plot+Position)** summary:

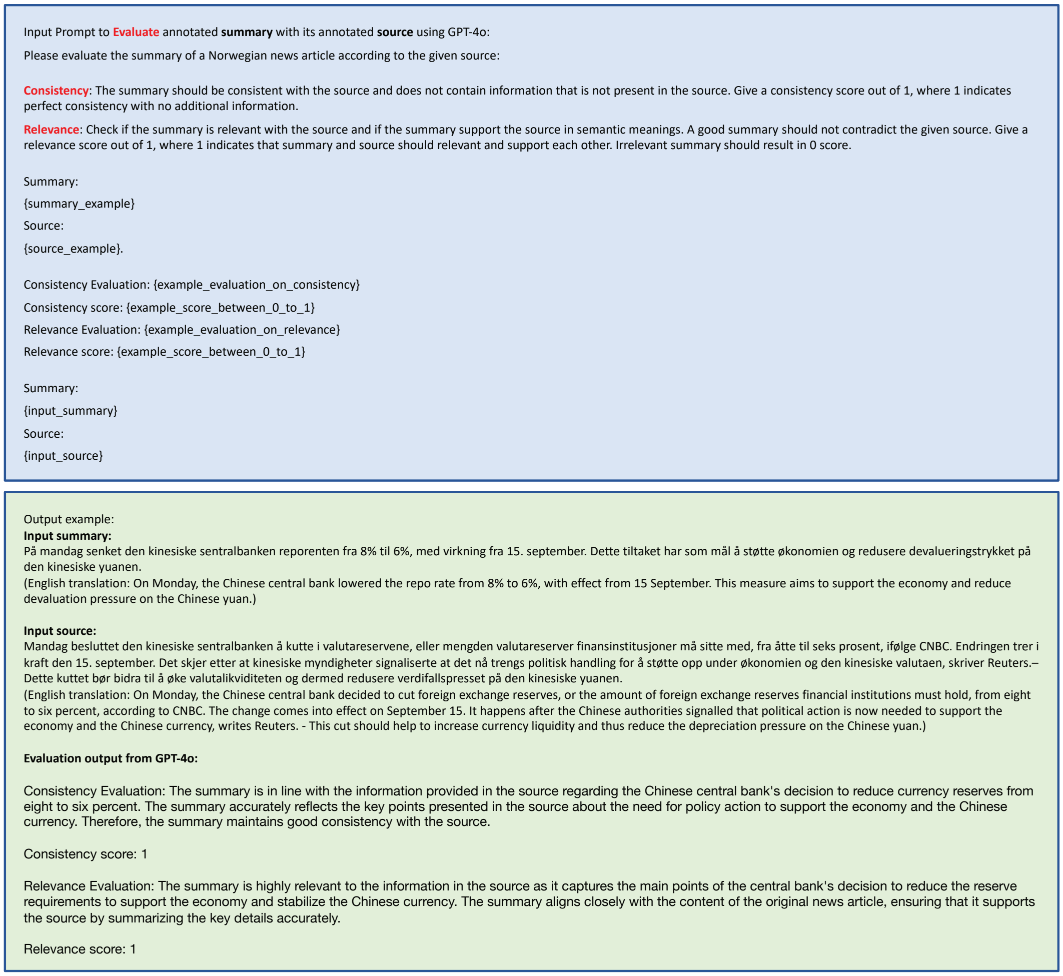
Du er en assistent som hjelper brukere med å skrive personlige sammendrag av gitte nyhetsartikler basert på deres historiske skrivestil og preferanser. Gitt en nyhetsartikkel og eksempler på brukerens tidligere sammendrag, generer et personlig tilpasset sammendrag på to til tre setninger som stemmer overens med brukerens unike stemme og stil.

Eksempel 0:
Artikkel: {historical_annotated_article0}
Sammendrag: {worker_summary0}
Entitetspreferanser: {entites_from_worker_summary_and_source0}
Tomtpreferanser: {plot_from_worker_summary_and_source0}
Stillingspreferanser: {position_from_worker_summary_and_source0}

Eksempel 1:
Artikkel: {historical_annotated_article1}
Sammendrag: {worker_summary1}
Entitetspreferanser: {entites_from_worker_summary_and_source1}
Tomtpreferanser: {plot_from_worker_summary_and_source1}
Stillingspreferanser: {position_from_worker_summary_and_source1}

Artikkel: {input_article}
Sammendrag:

Figure 4: Prompt for generating personalized summaries with all factors.



Input Prompt to **Evaluate** annotated **summary** with its annotated **source** using GPT-4o:

Please evaluate the summary of a Norwegian news article according to the given source:

Consistency: The summary should be consistent with the source and does not contain information that is not present in the source. Give a consistency score out of 1, where 1 indicates perfect consistency with no additional information.

Relevance: Check if the summary is relevant with the source and if the summary support the source in semantic meanings. A good summary should not contradict the given source. Give a relevance score out of 1, where 1 indicates that summary and source should relevant and support each other. Irrelevant summary should result in 0 score.

Summary:
{summary_example}
Source:
{source_example}.

Consistency Evaluation: {example_evaluation_on_consistency}
Consistency score: {example_score_between_0_to_1}
Relevance Evaluation: {example_evaluation_on_relevance}
Relevance score: {example_score_between_0_to_1}

Summary:
{input_summary}
Source:
{input_source}

Output example:
Input summary:
På mandag senket den kinesiske sentralbanken reporenten fra 8% til 6%, med virkning fra 15. september. Dette tiltaket har som mål å støtte økonomien og redusere devalueringstrykket på den kinesiske yuanen.
(English translation: On Monday, the Chinese central bank lowered the repo rate from 8% to 6%, with effect from 15 September. This measure aims to support the economy and reduce devaluation pressure on the Chinese yuan.)

Input source:
Mandag besluttet den kinesiske sentralbanken å kutte i valutareservene, eller mengden valutareserver finansinstitusjoner må sitte med, fra åtte til seks prosent, ifølge CNBC. Endringen trer i kraft den 15. september. Det skjer etter at kinesiske myndigheter signaliserte at det nå trengs politisk handling for å støtte opp under økonomien og den kinesiske valutaen, skriver Reuters.– Dette kuttet bør bidra til å øke valutalikviditeten og dermed redusere verdifallspresset på den kinesiske yuanen.
(English translation: On Monday, the Chinese central bank decided to cut foreign exchange reserves, or the amount of foreign exchange reserves financial institutions must hold, from eight to six percent, according to CNBC. The change comes into effect on September 15. It happens after the Chinese authorities signalled that political action is now needed to support the economy and the Chinese currency, writes Reuters. - This cut should help to increase currency liquidity and thus reduce the depreciation pressure on the Chinese yuan.)

Evaluation output from GPT-4o:

Consistency Evaluation: The summary is in line with the information provided in the source regarding the Chinese central bank's decision to reduce currency reserves from eight to six percent. The summary accurately reflects the key points presented in the source about the need for policy action to support the economy and the Chinese currency. Therefore, the summary maintains good consistency with the source.

Consistency score: 1

Relevance Evaluation: The summary is highly relevant to the information in the source as it captures the main points of the central bank's decision to reduce the reserve requirements to support the economy and stabilize the Chinese currency. The summary aligns closely with the content of the original news article, ensuring that it supports the source by summarizing the key details accurately.

Relevance score: 1

Figure 5: Prompt for evaluating annotated summary w.r.t. its source using GPT-4o and an example output.

A.3 Supplementary results for PersonalSum

Table 1: 5-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	37.53 ±14.74	**38.10** ±14.88	37.62 ±15.19	37.32 ±16.04	37.73 ±15.73	37.32 ±15.28	37.62 ±15.45	37.23 ±15.76	37.66 ±15.44
	Rouge-2	16.83 ±12.96	**17.40** ±13.11	17.10 ±13.10	17.11 ±13.56	17.29 ±13.61	16.76 ±13.25	17.04 ±13.42	16.83 ±13.51	17.22 ±13.47
	Rouge-L	26.77 ±13.26	27.08 ±13.27	26.70 ±13.48	26.82 ±14.06	**27.10** ±13.95	26.47 ±13.55	26.79 ±13.62	26.64 ±13.97	26.94 ±13.66
	BERTScore	74.95 ±5.35	**75.10** ±5.32	74.79 ±5.50	74.71 ±6.07	74.96 ±5.92	74.71 ±5.67	74.88 ±5.66	74.80 ±5.87	74.87 ±5.73
Gemini 1.0 Pro	Rouge-1	32.27 ±13.32	33.50 ±13.63	33.74 ±13.64	32.72 ±13.48	**33.91** ±13.65	33.89 ±13.14	33.85 ±13.56	33.66 ±13.59	33.64 ±13.17
	Rouge-2	12.35 ±10.65	13.18 ±11.08	13.29 ±10.73	12.51 ±10.83	13.46 ±11.57	13.06 ±10.84	**13.61** ±11.03	13.54 ±11.38	13.14 ±10.64
	Rouge-L	22.94 ±11.62	23.89 ±11.99	23.96 ±11.92	23.38 ±11.87	**24.36** ±12.09	23.99 ±11.77	23.93 ±11.98	24.10 ±12.10	23.95 ±11.55
	BERTScore	73.53 ±5.14	73.95 ±5.30	73.94 ±5.15	73.64 ±5.00	**73.98** ±5.27	73.90 ±5.06	73.88 ±5.10	73.95 ±5.13	73.93 ±5.13
NorwAI-Mixtral-8x7B-instruct	Rouge-1	34.16 ±12.99	**34.32** ±13.56	33.79 ±13.67	34.21 ±13.63	33.93 ±13.80	33.96 ±13.99	34.08 ±13.85	34.15 ±13.88	33.75 ±13.99
	Rouge-2	13.42 ±10.50	13.86 ±10.98	13.66 ±11.11	13.87 ±11.13	13.81 ±11.17	**13.98** ±11.30	13.73 ±11.21	13.78 ±11.21	13.90 ±11.38
	Rouge-L	23.64 ±10.81	**24.33** ±11.47	24.11 ±11.77	24.33 ±11.67	24.09 ±11.57	24.27 ±11.94	24.29 ±11.80	24.26 ±11.78	24.21 ±11.93
	BERTScore	73.51 ±4.70	73.90 ±4.97	73.86 ±5.05	73.90 ±5.04	73.90 ±5.06	73.89 ±5.14	**73.93** ±5.03	73.86 ±5.12	73.89 ±5.15

Table 2: 10-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	37.40 ±15.28	37.41 ±15.16	37.00 ±15.47	36.63 ±16.22	36.62 ±16.01	37.33 ±15.68	37.36 ±15.79	**37.64** ±15.76	37.10 ±15.97
	Rouge-2	17.00 ±13.45	16.95 ±13.20	17.02 ±13.31	16.78 ±13.54	16.62 ±13.60	16.99 ±13.13	17.29 ±13.70	**17.34** ±13.69	17.01 ±13.36
	Rouge-L	26.94 ±13.69	26.59 ±13.35	26.43 ±13.64	26.14 ±14.08	26.22 ±14.05	26.64 ±13.59	26.87 ±14.23	**27.06** ±13.93	26.60 ±13.89
	BERTScore	**74.99** ±5.57	74.84 ±5.38	74.66 ±5.61	74.46 ±5.95	74.51 ±6.05	74.70 ±5.72	74.86 ±5.75	74.92 ±5.77	74.66 ±5.93
Gemini 1.0 Pro	Rouge-1	33.22 ±13.22	34.79 ±14.11	34.17 ±13.79	**35.78** ±14.21	34.82 ±14.11	35.01 ±13.78	34.46 ±14.26	34.70 ±14.56	34.63 ±14.37
	Rouge-2	12.82 ±10.22	14.39 ±11.16	13.56 ±11.41	**14.55** ±11.81	14.05 ±11.34	14.43 ±11.12	13.86 ±11.47	14.25 ±11.65	13.86 ±11.36
	Rouge-L	23.06 ±11.22	24.65 ±11.82	23.83 ±11.90	**24.93** ±11.84	23.95 ±11.51	24.43 ±11.61	23.70 ±11.68	24.38 ±12.06	24.05 ±11.60
	BERTScore	73.97 ±4.62	74.26 ±4.74	74.00 ±4.79	**74.56** ±4.63	74.33 ±4.69	74.37 ±4.79	74.04 ±4.99	74.25 ±4.92	74.21 ±4.84
NorwAI-Mixtral-8x7B-instruct	Rouge-1	32.97 ±12.09	33.12 ±11.89	33.32 ±12.45	33.67 ±12.36	33.21 ±12.80	33.46 ±12.81	33.18 ±12.25	**34.05** ±12.25	33.75 ±12.79
	Rouge-2	12.67 ±11.30	13.07 ±10.89	13.22 ±11.10	13.46 ±11.03	13.11 ±11.17	13.38 ±10.60	13.29 ±10.36	**13.67** ±10.71	13.60 ±10.35
	Rouge-L	22.75 ±10.71	23.03 ±10.45	23.16 ±10.96	23.52 ±11.10	23.39 ±10.76	23.46 ±11.24	23.37 ±11.30	**23.94** ±10.72	23.83 ±11.71
	BERTScore	73.26 ±4.56	73.39 ±4.49	73.51 ±5.03	73.78 ±4.86	73.62 ±5.13	73.69 ±4.74	73.57 ±5.01	**73.84** ±4.82	73.59 ±5.20

A.4 Supplementary results for 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	39.42 ±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	18.22 ±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	28.06 ±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	75.73 ±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	36.46 ±14.10	36.32 ±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	15.11 ±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	25.76 ±11.89
	BERTScore	74.45 ±4.95	74.82 ±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	33.87 ±12.14	33.20 ±11.85	33.23 ±13.19	33.65 ±12.49	33.19 ±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	13.35 ±10.21
	Rouge-L	22.66 ±10.26	23.62 ±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	73.77 ±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

A.5 Error Case Analysis on PersonalSum

Figure 6 shows the improvements for each worker using personalized prompt methods compared to prompting for generating generic summaries. We can observe that some workers, such as workers 3 and 12, show significant improvements with personalized prompting, while others, such as workers 1 and 19, exhibit the opposite trend. To understand the reason for the performance degradation, we selected a case from worker 1 for analysis, as shown in Figure 9. We found that the current article lacks entities related to the worker's previous annotations, and the position information also differs. However, the worker is interested in the storyline and event cause, which has been captured by GPT-3.5-Turbo. The worker's interest is concentrated at the beginning of the article, while the generated summary covers most of it. This issue also appears in other error cases, making it difficult to capture the user's historical interests. One reason is that users cannot choose the articles they want to annotate. This limitation motivated our design to collect Topic-centric PersonalSum.

A.6 Statistics for Topic-centric PersonalSum

Topic-centric PersonalSum includes a total of 276 personalized summaries from 72 news articles annotated by 14 distinct Amazon Turkers. Specifically, Figure 7 shows the basic statistics of annotators, news articles and summaries for Topic-centric PersonalSum. Figure 8 shows the distributions of sources and average word counts per summary for machine-generated and human-annotated personalized summaries in Topic-centric PersonalSum.

Table 4: 10-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	37.36 ±14.06	37.74 ±14.30	37.99 ±13.87	37.96 ±15.51	38.39 ±13.85	37.53 ±15.69	36.80 ±15.29	**38.89** ±13.94	38.68 ±14.23
	Rouge-2	16.70 ±11.91	17.25 ±12.46	17.65 ±11.69	17.68 ±12.37	17.58 ±11.60	17.57 ±12.83	17.71 ±12.77	**18.01** ±11.58	17.97 ±12.27
	Rouge-L	26.67 ±12.26	27.19 ±13.10	27.09 ±12.25	26.72 ±12.86	27.18 ±12.16	26.89 ±13.30	27.48 ±13.85	27.63 ±11.86	**28.07** ±12.60
	BERTScore	74.90 ±5.12	75.17 ±5.26	75.09 ±5.04	74.93 ±5.64	75.16 ±5.42	75.31 ±5.30	75.20 ±5.25	75.50 ±4.91	**75.57** ±5.02
Gemini 1.0 Pro	Rouge-1	35.94 ±12.65	37.45 ±14.01	**38.44** ±13.10	37.08 ±14.43	37.52 ±13.50	37.32 ±13.88	36.89 ±13.53	37.14 ±14.27	37.90 ±13.51
	Rouge-2	14.60 ±10.45	16.05 ±11.57	16.33 ±10.77	15.52 ±11.36	16.16 ±11.27	16.20 ±11.33	15.76 ±10.93	**16.51** ±12.30	16.24 ±11.37
	Rouge-L	24.87 ±10.85	26.76 ±12.67	26.55 ±11.49	25.85 ±12.08	26.24 ±12.43	26.27 ±12.72	26.12 ±12.25	**26.91** ±13.47	26.40 ±12.13
	BERTScore	74.74 ±4.96	74.97 ±5.20	75.04 ±4.99	74.65 ±5.63	74.92 ±5.19	74.79 ±5.52	74.76 ±5.16	75.14 ±5.64	**75.18** ±5.24
NorwAI-Mixtral-8x7B-instruct	Rouge-1	34.00 ±12.94	35.78 ±13.36	34.93 ±13.58	34.81 ±13.53	35.15 ±13.05	**36.00** ±13.94	35.03 ±13.90	34.96 ±13.48	35.14 ±13.64
	Rouge-2	13.33 ±9.96	15.13 ±9.80	14.28 ±10.30	14.17 ±10.21	14.41 ±9.76	**15.50** ±10.54	14.39 ±10.96	14.67 ±10.19	14.51 ±10.43
	Rouge-L	23.93 ±10.85	24.91 ±10.57	24.78 ±11.14	24.22 ±10.93	24.44 ±10.42	**25.38** ±11.46	24.78 ±11.72	24.64 ±11.13	24.96 ±11.33
	BERTScore	73.63 ±4.86	74.38 ±4.93	74.11 ±4.89	74.19 ±5.03	74.30 ±5.03	**74.43** ±5.16	74.29 ±5.27	74.24 ±5.11	74.23 ±5.07

Table 5: Entailment scores of different LLMs on PersonalSum. Best results are on bold.

	Models	Generic	Direct	Entity	Plot	Position	Entity+Plot	Entity+Position	Plot+Position	All
PersonalSum Dataset										
2-shot	GPT-3.5 Turbo	94.91	94.22	95.10	94.81	94.03	94.61	94.81	**95.69**	93.24
	Gemini 1.0 Pro	90.04	90.26	**91.56**	90.15	86.47	90.80	88.64	87.77	89.07
	NorwAI-Mixtral	93.14	95.00	97.16	96.87	95.40	**97.26**	96.47	96.57	95.10
5-shot	GPT-3.5 Turbo	95.17	94.07	**96.05**	94.95	95.17	**96.05**	94.07	94.95	94.95
	Gemini 1.0 Pro	90.80	90.68	**91.02**	88.96	90.26	88.64	90.15	89.61	89.61
	NorwAI-Mixtral	93.70	94.33	96.53	96.22	96.53	97.37	96.01	**97.58**	96.11
10-shot	GPT-3.5 Turbo	94.39	95.01	95.51	**96.01**	94.89	94.01	95.76	95.76	95.14
	Gemini 1.0 Pro	88.16	92.52	92.64	91.77	91.90	93.52	91.65	92.39	**93.76**
	NorwAI-Mixtral	94.28	95.08	**97.83**	96.00	95.64	96.80	97.48	96.22	97.14
Topic-centric PersonalSum Dataset										
2-shot	GPT-3.5 Turbo	91.50	92.71	88.66	90.69	92.31	89.88	90.28	**93.12**	91.90
	Gemini 1.0 Pro	**95.15**	92.07	91.19	92.51	89.43	92.07	92.95	87.22	92.51
	NorwAI-Mixtral	90.28	93.93	96.76	97.98	94.74	95.14	94.33	**98.38**	95.95
5-shot	GPT-3.5 Turbo	91.94	93.84	92.89	91.00	91.00	**95.73**	93.35	91.47	92.42
	Gemini 1.0 Pro	93.84	92.89	91.47	93.36	**94.79**	**94.79**	93.84	92.89	90.05
	NorwAI-Mixtral	91.24	95.39	95.85	94.47	94.01	**96.77**	94.93	96.31	94.93
10-shot	GPT-3.5 Turbo	91.30	95.03	93.17	**96.27**	93.79	92.55	94.41	95.03	93.79
	Gemini 1.0 Pro	**94.70**	91.39	93.38	**94.70**	91.39	94.04	90.73	88.74	88.74
	NorwAI-Mixtral	93.60	90.12	**97.09**	93.60	94.77	95.35	96.51	94.19	**97.09**

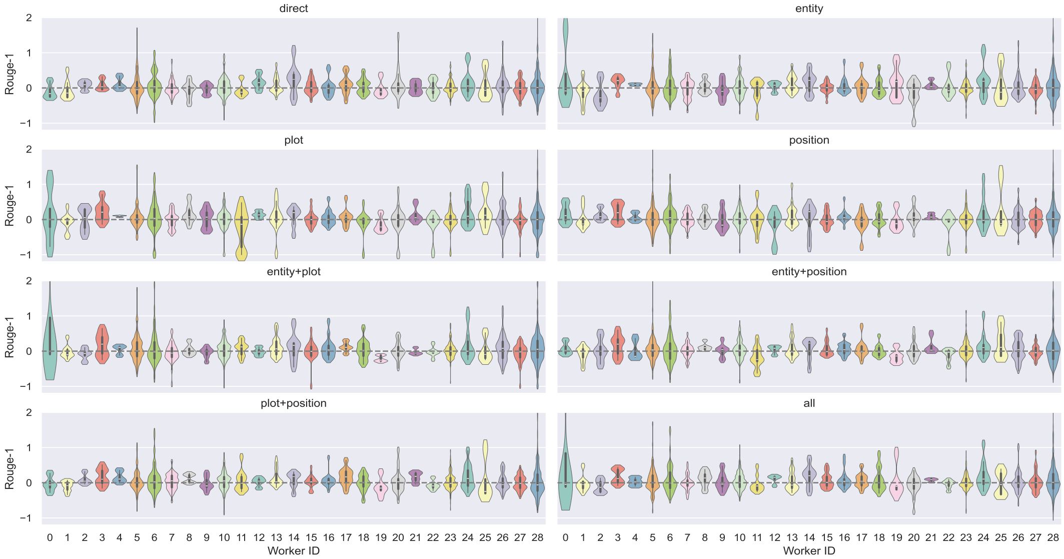
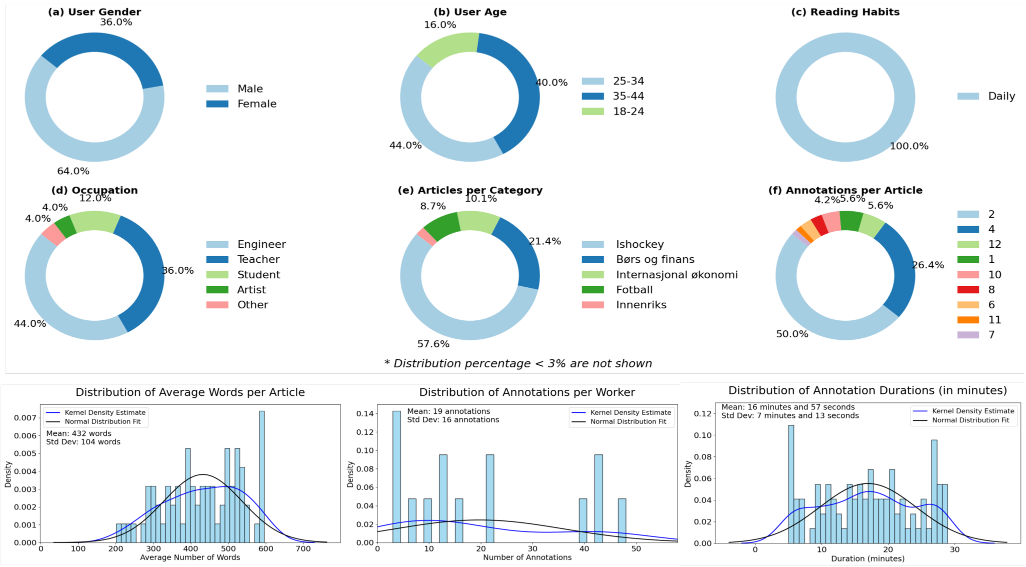
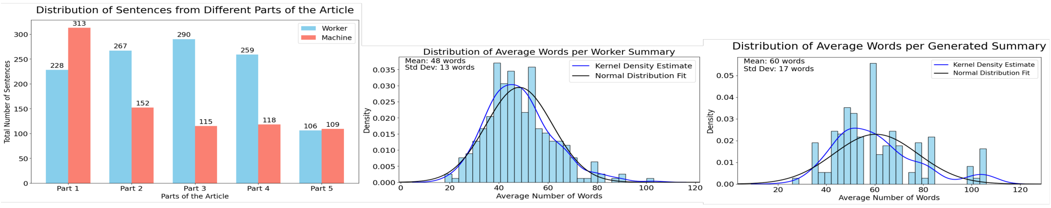


Figure 6: Violin plot showing improvements in the ROUGE-1 score from personalized prompting compared to generic summaries using GPT-3.5 Turbo for each worker. The X-axis represents worker IDs, and the Y-axis represents the ROUGE-1 score improvements for PersonalSum dataset.



(g) (h) (i)

Figure 7: Statistics of the annotators, news articles and summaries for Topic-centric PersonalSum.



(a) (b)

Figure 8: (a) The distribution of sources of machine-generated summaries and human-annotated personalized summaries in the Topic-centric PersonalSum. (b) The distribution of average words per machine-generated summary and human-annotated summary in the Topic-centric PersonalSum.

Extracted factors of Entity, Plot and Position in worker's 5-shot summaries:

Input 1:

Entity: japanske, torsdag, sentralbanken, 0, 1 prosentpoeng, kinesiske, natten, reporente, Kina, Japan, Reuters

Plot: Event storyline

Position: [0, 0, 1, 0, 0]

Input 2:

Entity: devalueringstrykket, sentralbanken, reprinsen, kinesiske, kinesiske yuanen, september, økonomien, finansinstitusjoner, CNBC, valutareserver, kinesiske sentralbanken, Reuters, verdifallspresset, valutalikviditeten

Plot: Event storyline, Event Cause

position: [0, 0, 1, 1, 0]

Input 3:

Entity: seks måneder, Grete, Staffordshire bullterrieren, skateboard, Barna, Oslo, skateparker, Daniel Engen, Staffordshire bull terrieren, Aftenposten Junior

Plot: Event storyline, Event Cause

Position: [2, 3, 2, 0, 0]

Input 4:

Entity: Venstre, Miljøpartiet De Grønne, Noah, Senterpartiet

Plot: Event storyline, Event Cause
Position: [0, 3, 2, 0, 0]

Input 5:

Entity: Sverige, Hilda, Marika Andersson, Kua Hilda

Plot: Event storyline, Event Cause, Event result

Position: [4, 4, 5, 3, 0]

Golden worker summary and its extracted factors of Entity, Plot and Position:

Romanen "Han, Solo" følger Robins hjertesorg etter at kjæresten hans, ved navn Han, avslutter forholdet deres. Tittelen antyder karakteren Han Solo fra «Star Wars», noe som gjenspeiler Robins interesse for franchisen og Hans forkjærlighet for solo- drikking.

(English translation: The novel "Han, Solo" follows Robin's heartbreak after his lover, named Han, ends their relationship. The title suggests the character Han Solo from 'Star Wars', reflecting Robin's interest in the franchise and Han's penchant for solo drinking.)

Entity: Robins, Star Wars, Robin, Cappelen Damm, Han, Han Solo

Plot: Event storyline, Event Cause

Position: [3, 2, 0, 0, 0]

Generated summary with Plot factor using GPT-3.5-Turbo:

"Han, solo" er en roman som følger Robin gjennom en periode med kjærlighetssorg etter at kjæresten gjør det slutt. Boken tar leseren med på en reise gjennom Robins liv, fra dagliglivet i Bergen under pandemien til hans bakgrunn fra Flandern i Belgia. Selv om boken har en tendens til å være selvopptatt og detaljert, inneholder den effektive skildringer av ensomhet, refleksjoner over kjærlighet og teknologiens innvirkning på relasjoner.

(English translation: "Han, solo" is a novel that follows Robin through a period of heartbreak after his girlfriend breaks up with him. The book takes the reader on a journey through Robin's life, from daily life in Bergen during the pandemic to his background from Flanders in Belgium. Although the book tends to be self-absorbed and detailed, it contains effective depictions of loneliness, reflections on love and the impact of technology on relationships.)

Extracted Plot using GPT-4o, where <c>, <s> and <e> represent Event cause, Event storyline and Event result, respectively. English translations are in the bracket:

<c> I byrjinga av romanen gjer Robins kjærast det slutt med han, noko som utløyser ei stor kjærleikssorg. </c>

(At the beginning of the novel, Robin's lover breaks up with him, which triggers a great heartbreak.)

<s> «Han, solo» er ein roman som handlar om Robin. </s>

(«Han, solo» is a novel about Robin.)

<s> Ekskjærasten blir gitt namnet Han, som både understrekar at dette er ein mann og hintar til karakteren Han Solo frå «Star Wars». </s>

(The ex-boyfriend is given the name Han, which both emphasizes that this is a man and hints at the character Han Solo from «Star Wars».)

<s> Men det dukkar opp i teksten at Robin er opptatt av «Star Wars» og at Han likte å drikke solo. </s>

(But it appears in the text that Robin is interested in "Star Wars" and that Han liked to drink solo.)

....

<s> Den er svært tett på hovudpersonen Robin, som deler namn og bakgrunn med sin forfattar, og fortel både om daglegliv og biblioteksjobb i Bergen under pandemien og om Robins bakgrunn frå Flandern i Belgia, studiar i Amsterdam og Oslo. </s>

(It is very close to the main character Robin, who shares a name and background with its author, and tells both about daily life and library work in Bergen during the pandemic and about Robin's background from Flanders in Belgium, studies in Amsterdam and Oslo.)

......

<s> Men forsterka av pandemiens isolasjon, går Robin etter brotet gjennom si eiga kjensle av gjentakande einsemd, han jamfører med litteratur og film, og skaper då gode tablå og refleksjonar som er slåande, kanskje særleg fordi forteljestilen er enkel og nedpå, med svært kvardagslege gjeremål og fullstendig blotta for patos. </s>

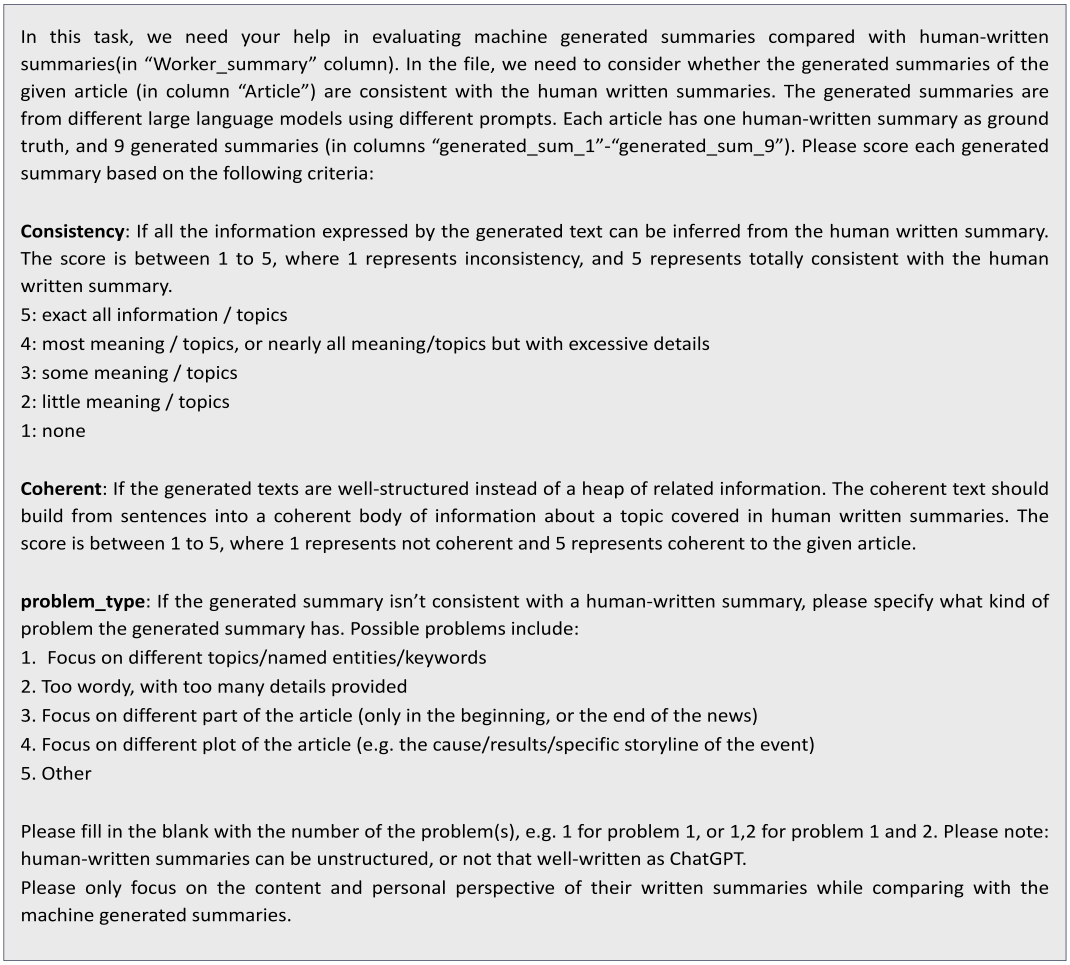
(But strengthened by the isolation of the pandemic, after the break Robin goes through his own feeling of recurring loneliness, he compares literature and film, and then creates good tableaus and reflections that are striking, perhaps especially because the narrative style is simple and to the point, with very everyday tasks and completely devoid of pathos.)

......

<e> Denne utgjevinga er ujamn, men på sitt beste særs lovande. </e>

(This release is uneven, but at best very promising.)

Figure 9: The blue box shows the entity, plot, and position factors extracted from the user's five most recent annotation histories. The green box displays the user's current annotation summary, the corresponding extracted factors, and a generated result from GPT-3.5-Turbo with the plot factor. The yellow box shows part of the plot components extracted by GPT-4o related to the generated summary and the worker's annotated summary. (The corresponding English translation is in brackets.)



In this task, we need your help in evaluating machine generated summaries compared with human-written summaries(in “Worker_summary” column). In the file, we need to consider whether the generated summaries of the given article (in column “Article”) are consistent with the human written summaries. The generated summaries are from different large language models using different prompts. Each article has one human-written summary as ground truth, and 9 generated summaries (in columns “generated_sum_1”-“generated_sum_9”). Please score each generated summary based on the following criteria:

Consistency: If all the information expressed by the generated text can be inferred from the human written summary. The score is between 1 to 5, where 1 represents inconsistency, and 5 represents totally consistent with the human written summary.
5: exact all information / topics
4: most meaning / topics, or nearly all meaning/topics but with excessive details
3: some meaning / topics
2: little meaning / topics
1: none

Coherent: If the generated texts are well-structured instead of a heap of related information. The coherent text should build from sentences into a coherent body of information about a topic covered in human written summaries. The score is between 1 to 5, where 1 represents not coherent and 5 represents coherent to the given article.

problem_type: If the generated summary isn’t consistent with a human-written summary, please specify what kind of problem the generated summary has. Possible problems include:

1. Focus on different topics/named entities/keywords
2. Too wordy, with too many details provided
3. Focus on different part of the article (only in the beginning, or the end of the news)
4. Focus on different plot of the article (e.g. the cause/results/specific storyline of the event)
5. Other

Please fill in the blank with the number of the problem(s), e.g. 1 for problem 1, or 1,2 for problem 1 and 2. Please note: human-written summaries can be unstructured, or not that well-written as ChatGPT.
Please only focus on the content and personal perspective of their written summaries while comparing with the machine generated summaries.

Figure 10: Instructions for the human evaluation of personalized summaries generated by LLMs.

B Datasheets for PersonalSum dataset

B.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
The dataset was created to introduce a personalized summarization dataset for benchmarking and evaluating existing LLMs. Unlike prior open-source summarization datasets, PersonalSum focuses on collecting personalized summaries from public perspectives. It includes various attributes such as sources, multiple document-grounded question answers for each article, and human-corrected machine-generated generic summaries. These features are critical for advancing the field of personalized generation tasks using LLMs.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created by Lemei Zhang, Peng Liu, Marcus Tiedemann Oekland Henriksboe, Even W. Lauvrak, Jon Atle Gulla, and Heri Ramampiaro at Norwegian University of Science and Technology.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
PersonalSum project has been funded by the SFI NorwAI, (Centre for Research-based Innovation, 309834), the Research Council of Norway and the partners of the SFI NorwAI.

Any other comments?
None.

B.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
The instances are Norwegian news articles which include different attributes:

- Human-corrected machine-generated summary with source;
- Personalized summaries annotated by distinct annotators with source;
- Multiples question answer pairs with source.

How many instances are there in total (of each type, if appropriate)?
It contains 1099 personalized summaries from 441 news articles annotated by 39 distinct Amazon Turkers (See Section 3.2) in original PersonalSum, and 276 personalized summaries for 72 article in Topic-centric PersonalSum (See Section 5).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The dataset is a sample of instances. It is intended to be a random sample of Norwegian news articles from 10 different news categories. The annotators of personalized summaries are carefully selected through qualification tests to ensure that they cover different age groups and are with different backgrounds. During the annotation process, we specifically targeted annotators from underrepresented age groups based on age distribution. However, due to the scarcity of annotators in the Norwegian market and the stringent quality control of the annotated summaries, the final group of annotators was not evenly distributed. Nonetheless, according to statistics of the proposed dataset in Section 3.2 and Section 5, the annotators covered the age range of 18-44 years from the general public.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance consists of the textual news content associated with category, machine-generated summary with annotated source, personalized summaries with annotated sources, and multiple document-grounded question answer pairs with annotated sources. All labeled data passes predefined quality control rules described in Section 3.1.

Is there a label or target associated with each instance? If so, please provide a description.
Yes. For document-grounded question answering, the label is the annotated free-form textual response to the question based on the context of the input news article. For personalized summarization, the label is the annotated free-form textual summary of the input news article provided by a specific annotator. For the explainable generation task, the label consists of sentences from the news article that are associated with the generated summary or the answer to a question related to the news article.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
A few annotators chose not to disclose their age range; thus, this attribute is recorded as "prefer not to say" due to privacy protection purpose.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? s made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
Each annotators have at least three summaries for different news articles. Thus, annotators' historical records can be extracted.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
No. Given the limited data volume, we employed the entire dataset for in-context learning.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
The dataset includes crowdsourced input from a low-resource language group, making quality control challenging. We cannot filter annotators based on their successful annotation rate, as this would limit the pool and reduce data diversity. To address this, we iteratively monitor quality through manual inspection and LLM assistance. Despite these efforts, we cannot ensure the data is entirely error-free due to its subjective nature. For example, some annotators provided only partial sources associated with their annotated summaries, but we consider these valid if they contain the main information of the summary.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is entirely self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' nonpublic communications)? If so, please provide a description.
Some news articles contain copyright information but have been authorized by the copyright holder. The dataset is used solely for research purposes under the CC BY-NC 4.0 license.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
Age and gender information of the annotators were collected for statistical analysis but are not included in the dataset, eliminating the risk of privacy leakage. User IDs are anonymized to prevent the identification of specific individuals.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
Some personal information is retained as described in previous questions. However, since this dataset is dedicated to collecting users' personalized focus for given news articles, we anonymize user IDs to ensure that specific users cannot be identified.

Does the dataset contain data that might be considered sensitive in anyway (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
No.

Any other comments?
No.

B.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Some news articles contain copyright information from our partner Schibsted[4]. The use and release of the dataset have been approved by the data provider.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
The news articles are shared by data providers. Personalized summaries are collected via Amazon Turk. Other annotations including corrected generic summaries and document-grounded QA pairs are collected via recruited colleague students and OpenAI gpt-4[5].

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The sample of instances collected is Norwegian news articles provided by media companies. The sample is arbitrarily selected from 10 news categories including 465 samples.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
The collection process involves three stages detailed in Section 3.1. In the first stage, we recruited 3 colleague students from the department of computer science at Norwegian University of Science and Technology. They are compensated 230 NOK (approx. $21,75 USD) per hour. In the second stage, the annotators are crowdworkers recruited from Amazon Turk, who are compensated $6 USD per approved HIT, reflecting an hourly rate of $18 USD.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
The data in the first stage was collected from August 2023 to December 2023, while the second and third stages were iteratively conducted from January 2024 to April 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
Not applicable to this project.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
In the first stage, data collection involved revising machine-generated summaries and adding sources, accomplished by directly engaging recruited 3 college students. In the second phase, data was collected via Amazon Turk, a crowdsourcing platform.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Yes. In the first stage, the data is collected through a self-developed platform shown in Figure 11, while in the second stage, the data is collected using Amazon Turk shown in Figure 12 with the code available on our GitHub repo[6].

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
The sensitive data is only used for statistics, and Amazon Turk is a crowdsourcing platform for data collection.

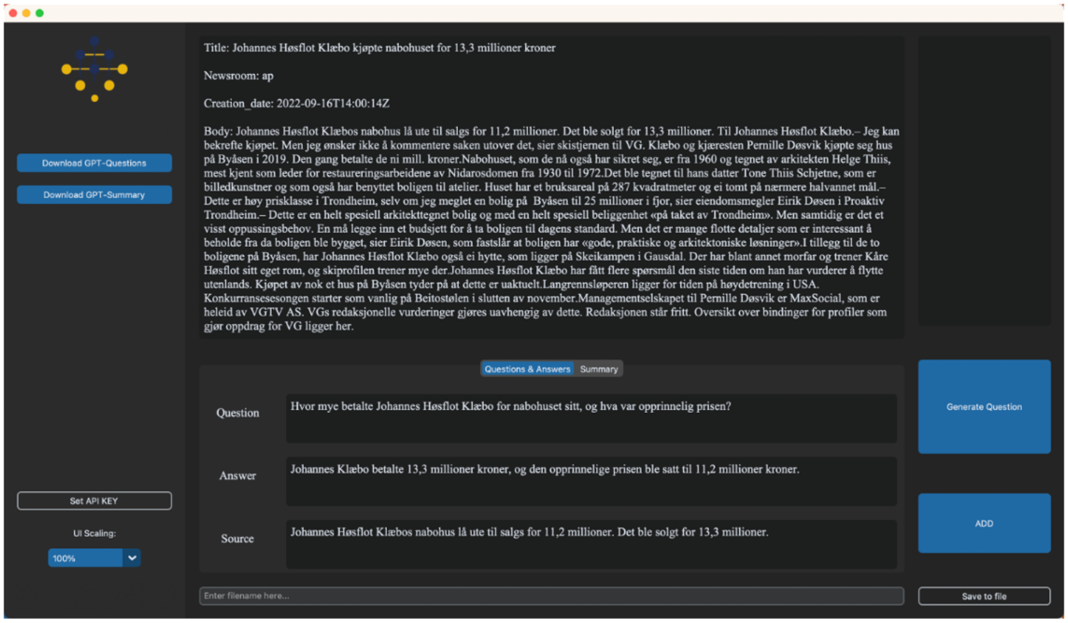
If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link

---

[4] `https://schibsted.com/`

[5] `https://openai.com/api/pricing/`

[6] `https://github.com/SmartmediaAI/PersonalSum`



Title: Johannes Høsflot Klæbo kjøpte nabohuset for 13,3 millioner kroner

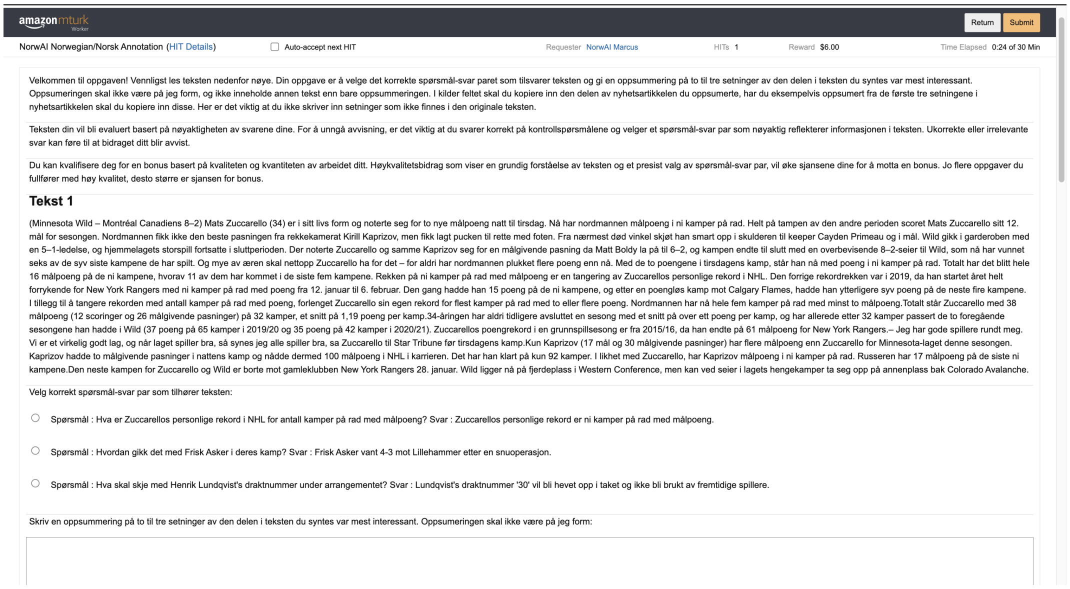
Newsroom: ap

Creation_date: 2022-09-16T14:00:14Z

Body: Johannes Høsflot Klæbos nabohus lå ute til salgs for 11,2 millioner. Det ble solgt for 13,3 millioner. Til Johannes Høsflot Klæbo.– Jeg kan bekrefte kjøpet. Men jeg ønsker ikke å kommentere saken utover det, sier skistjernen til VG. Klæbo og kjæresten Pernille Døsvik kjøpte seg hus på Byåsen i 2019. Den gang betalte de ni mill. kroner.Nabohuset, som de nå også har sikret seg, er fra 1960 og tegnet av arkitekten Helge Thiis, mest kjent som leder for restaureringsarbeidene av Nidarosdomen fra 1930 til 1972.Det ble tegnet til hans datter Tone Thiis Schjetne, som er billedkunstner og som også har benyttet boligen til atelier. Huset har et bruksareal på 287 kvadratmeter og ei tomt på nærmere halvannet mål.– Dette er høy prisklasse i Trondheim, selv om jeg meglet en bolig på Byåsen til 25 millioner i fjor, sier eiendomsmegler Eirik Døsen i Proaktiv Trondheim.– Dette er en helt spesiell arkitekttegnet bolig og med en helt spesiell beliggenhet «på taket av Trondheim». Men samtidig er det et visst oppussingsbehov. En må legge inn et budsjett for å ta boligen til dagens standard. Men det er mange flotte detaljer som er interessant å beholde fra da boligen ble bygget, sier Eirik Døsen, som fastslår at boligen har «gode, praktiske og arkitektoniske løsninger».I tillegg til de to boligene på Byåsen, har Johannes Høsflot Klæbo også ei hytte, som ligger på Skeikampen i Gausdal. Der har blant annet morfar og trener Kåre Høsflot sitt eget rom, og skiprofilen trener mye der.Johannes Høsflot Klæbo har fått flere spørsmål den siste tiden om han har vurderer å flytte utenlands. Kjøpet av nok et hus på Byåsen tyder på at dette er uaktuelt.Langrennsløperen ligger for tiden på høydetrening i USA. Konkurransesesongen starter som vanlig på Beitostølen i slutten av november.Managementselskapet til Pernille Døsvik er MaxSocial, som er heleid av VGTV AS. VGs redaksjonelle vurderinger gjøres uavhengig av dette. Redaksjonen står fritt. Oversikt over bindinger for profiler som gjør oppdrag for VG ligger her.

Question	Hvor mye betalte Johannes Høsflot Klæbo for nabohuset sitt, og hva var opprinnelig prisen?
Answer	Johannes Klæbo betalte 13,3 millioner kroner, og den opprinnelige prisen ble satt til 11,2 millioner kroner.
Source	Johannes Høsflot Klæbos nabohus lå ute til salgs for 11,2 millioner. Det ble solgt for 13,3 millioner.

Figure 11: Data annotation platform used in stage 1.



NorwAI Norwegian/Norsk Annotation (HIT Details)

☐ Auto-accept next HIT

Requester NorwAI Marcus

HITs 1

Reward $6.00

Time Elapsed 0:24 of 30 Min

Velkommen 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 interessant. 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 setningene i 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. Ukorrekte eller irrelevante 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. Jo flere oppgaver du 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 Zuccarello sitt 12. 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 gikk i garderoben med en 5–1-ledelse, og hjemmelagets storspill fortsatte i sluttperioden. Der noterte Zuccarello og samme Kaprizov seg for en målgivende pasning da Matt Boldy la på til 6–2, og kampen endte til slutt med en overbevisende 8–2-seier til Wild, som nå har vunnet seks av de syv siste kampene de har spilt. Og mye av æren skal nettopp Zuccarello ha for det – for aldri har nordmannen plukket flere poeng enn nå. Med de to poengene i tirsdagens kamp, står han nå med poeng i ni kamper på rad. Totalt har det blitt hele 16 målpoeng på de ni kampene, hvorav 11 av dem har kommet i de siste fem kampene. Rekken på ni kamper på rad med målpoeng er en tangering av Zuccarellos personlige rekord i NHL. Den forrige rekordrekken var i 2019, da han startet året helt forrykende for New York Rangers med ni kamper på rad med poeng fra 12. januar til 6. februar. Den gang hadde han 15 poeng på de ni kampene, og etter en poengløs kamp mot Calgary Flames, hadde han ytterligere syv poeng på de neste fire kampene. I tillegg til å tangere rekorden med antall kamper på rad med poeng, forlenget Zuccarello sin egen rekord for flest kamper på rad med to eller flere poeng. Nordmannen har nå hele fem kamper på rad med minst to målpoeng.Totalt står Zuccarello med 38 målpoeng (12 scoringer og 26 målgivende pasninger) på 32 kamper, et snitt på 1,19 poeng per kamp.34-åringen har aldri tidligere avsluttet en sesong med et snitt på over ett poeng per kamp, og har allerede etter 32 kamper passert de to foregående 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 gode spillere rundt meg. Vi er et virkelig godt lag, og når laget spiller bra, så synes jeg alle spiller bra, sa Zuccarello til Star Tribune før tirsdagens kamp.Kun Kaprizov (17 mål og 30 målgivende pasninger) har flere målpoeng enn Zuccarello for Minnesota-laget denne sesongen. 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 målpoeng på de siste ni 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 Colorado Avalanche.

Velg korrekt spørsmål-svar par som tilhører teksten:

○ 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.

Skriv en oppsummering på to til tre setninger av den delen i teksten du syntes var mest interessant. Oppsumeringen skal ikke være på jeg form:

Figure 12: An example of the HIT on Amazon Turk.

or other access point to the mechanism (if appropriate).
N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
N/A

B.4 Preprocessing / cleaning / labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
Instances that did not pass quality control (See Section 3.1) were removed. Annotators and records with fewer than three valid annotations were excluded. Besides, we encrypted the worker ids to protect privacy.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
The dataset itself contains all the raw data. The sensitive data won't be shared.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
The code for quality control is available on Github.

Any other comments?
None.

B.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.
Only this paper at the time of publication.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
N/A

What (other) tasks could the dataset be used for?
The rich properties of PersonalSum enable its use in 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 public interests. It also facilitates comparative analyses of machine-generated and manually annotated summaries, showcasing the dataset's versatility and potential to advance research.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
The dataset can be used for research only. **Please** do not use it for any commercial purpose.

Are there tasks for which the dataset should not be used? If so, please provide a description.
No.

B.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
Yes. The dataset is publicly available on the internet.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is available both on GitHub `https://github.com/SmartmediaAI/PersonalSum` and Huggingface `https://huggingface.co/datasets/PersonalLab/PersonalSum`. The dataset does not have a DOI, and there is no other redundant archive.

When will the dataset be distributed? The dataset was first released in June 2024.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The license of this dataset is cc-by-nc-4.0. And there is a request to cite the corresponding paper if the dataset is used.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other
Unknown to authors of the datasheet.

B.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?
Lemei Zhang and Peng Liu are supporting/maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The curators of the dataset, Lemei Zhang and Peng Liu, can be contacted at `https://www.ntnu.edu/employees/lemei.zhang` and `https://www.ntnu.edu/employees/penl`, respectively.

Is there an erratum? If so, please provide a link or other access point.
Not for now. Any further updates will be uploaded to our GitHub repo and Huggingface.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?
This will be posted on the dataset repo.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
Not applicable.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
Not applicable.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
Please contact the builder of the dataset, cite the paper and follow the license.

Any other comments?
No.

C Responsibility Statement

The authors of PersonalSum bear all responsibility in case of any violation of rights or any other legal issues that arise from the use of this dataset. The information provided in the paper and the supplementary material is truthful and accurate. The authors have taken all possible measures to ensure the respect of privacy and ethical guidelines in the construction of this dataset.