Can Language Models Capture Human Writing Preferences for Domain-Specific Text Summarization?

Jingbao Luo[1], Ming Liu[2*], Ran Liu[3], Yongpan Sheng[4], Xin Hu[5], Gang Li[2], Peng Wu[6†],
[1] School of Cyber Science and Engineering, Nanjing University of Science and Technology
[2]School of Information Technology, Deakin University
[3]Institute of Information Engineering, Chinese Academy of Sciences
[4]College of Computer and Information Science, Southwest University
[5]School of Architecture and Built Environment, Deakin University
[6]School of Intelligent Manufacturing, Nanjing University of Science and Technology
{luojingbao, wupeng}@njust.edu.cn, liuran@iie.ac.cn, shengyp2011@gmail.com
{m.liu, xin.hu, gang.li}@deakin.edu.au

Abstract

With the popularity of large language models and their high-quality text generation capabilities, researchers are using them as auxiliary tools for text summary writing. Although summaries generated by these large language models are smooth and capture key information sufficiently, the quality of their output depends on the prompt, and the generated text is somewhat procedural to a certain extent. We construct LecSumm to verify whether language models truly capture human writing preferences, in which we recruit 200 college students to write summaries for lecture notes on ten different machine-learning topics and analyze writing preferences in real-world human summaries through the dimensions of length, content depth, tone & style, and summary format. We define the method of capturing human writing preferences by language models as fine-tuning pre-trained models with data and designing prompts to optimize the output of large language models. The results of translating the analyzed human writing preferences into prompts and conducting experiments show that both models still fail to capture human writing preferences effectively. Our LecSumm [1] dataset brings new challenges to finetuned and prompt-based large language models on the task of human-centered text summarization.

1 Introduction

With the vast amount of training data, the development of large language models (LLMs), such as the GPT series (OpenAI, 2024), the PaLM series (Aakanksha Chowdhery, 2022), Mistral (Jiang et al., 2023) and LLaMA (AI@Meta, 2024), have achieved remarkable success by unifying the generative paradigm with different NLP tasks(Wei et al., 2024a,c; Wan et al., 2023; Wang et al., 2023a; Qin et al., 2023; tse Huang et al., 2024). In specific NLP fields, such as text summarization, LLMs achieve decent performance without additional training data and even surpass traditional supervised fine-tuned models (Zhang et al., 2024). Recent studies employ LLMs as auxiliary tools for human-centered NLP (Passali et al., 2021; Hu et al., 2023), ranging from human-centered design to human-in-the-loop interaction with LLMs. When generating human-centered summaries with LLMs, specific human preferences can be incorporated through two different approaches : (i) Explicitly, add external constraints to the summarization model, such as prompt design and different hyperparameter settings. (ii) Implicitly, construct specific source-target summary datasets that reflect human preferences to finetune the language model, enabling it to learn the hidden preferences from the data. Our research question is: **Can language models really capture human writing preferences through prompting and in-context training?**

To answer this question, we first design a lecture note summarization task to discover human preferences in real-world data and construct a dataset containing human-centered summaries. The task framework is shown in Figure 1. We recruited 200 university students and designed a task of writing lecture note summaries: 10 different topics related to machine learning were given to the annotators, together with the corresponding lecture notes, and the participants were required to write summaries based on the lecture notes. There is no hard limitation (e.g., length, summary format) on the summary writing process; participants are allowed to use re-

*Corresponding author.
†Corresponding author.
[1]https://github.com/DeakinAINLP/LecSumm

Findings of the Association for Computational Linguistics: ACL 2025, pages 6073–6091

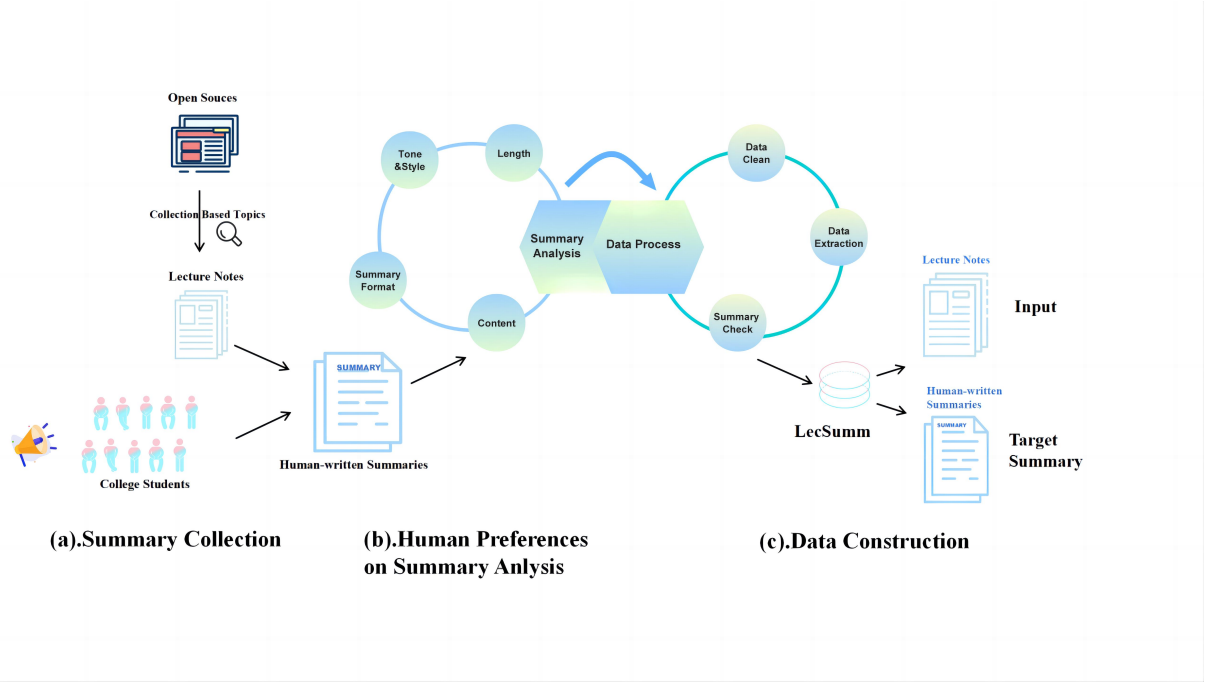

Figure 1: The framework of our work is divided into three stages: **(a) Summary Collection.** We designed a summary collection task and recruited annotators online. They were asked to write summaries based on the provided lecture notes gathered from open-source data on specific topics. **(b) Human Preferences Analysis on Summaries.** Human-written summaries were analyzed from four dimensions: length, structure, modality, and content depth to discover human needs. **(c) Dataset Construction.** The dataset was constructed after summary checking and data processing, with lecture notes as input sources and human-written summaries as target summaries.

lated materials to equip the lecture notes. The only limitation is that it has to be written by the participants and cannot be returned by machines. We observe that different annotators utilize a combination of various dimensions to reflect their writing preferences; these human preferences range in four dimensions: length, content, tongue&style, and summary format.

Then, we construct the LecSumm dataset, which includes the provided lecture notes and human-written summaries, and experiment with finetuned supervised models and prompt-based zero-shot LLMs. We find that language models can not capture human writing preferences well through model finetuned and prompt design. Our main contribution is presented as follows:

- We design a human-centered text summarization task to collect human-guided summaries for lecture notes and detect four writing preferences from human-guided summaries.
- We propose a LecSumm dataset containing human-centered summaries and verify the usability of the dataset through dataset analysis and model experiments.
- We analyze language models' capture ability of human writing preferences with finetuned and prompt design and find that language models can not capture human writing preferences well.

2 Lecture Note Summary Collection and Analysis

2.1 Human-centered Summary Collection

Summary Collection Task We collected machine learning lecture notes which cover ten major topics, as shown in Table 1, the lecture notes are all open source and can be found on the Internet, most of them are released by public universities. We recruited 200 university students from the IT department and asked them to write summaries for the ten topics after reading the lecture notes. The recruited people are required to have at least finished one machine learning related course, and each annotator is reimbursed with $100 to cover the annotation cost.

Annotator Statistic Recruited annotators are students from university IT departments. While recruiting student participants, we also asked the annotators to provide the following information: first language, qualification, and machine learning

	Topic
1	Machine Learning Overview
2	Data Wrangling
3	Clustering Algorithms
4	Principal Component Analysis (PCA)
5	A supervised learning algorithm
6	Linear regression
7	Support Vector Machine(SVM)
8	Decision tree algorithms
9	Ensemble learning
10	Neural Networks and Deep Learning

Table 1: These are ten topics covered by the machine learning lecture notes we collected.

Annotator	Percentage
First Language	82% Native English
Qualification	30% B.S / 70% M.S
ML Experience	65% experienced / 35% non-exp

Table 2: Annotator statistics

working experience. Table 2 shows that 82% of the annotators are native English speakers and 65% of them indicate that they have machine learning related working experience. It shows the high quality of the annotator group and will guarantee real human summaries in the annotation process.

2.2 Human Writing Preferences Detection on Summaries

Human writing preferences (McNamara et al., 2010) generally refer to the specific styles, expressions, or structures that individuals use in their writing. These preferences may be reflected in word choice, sentence length, grammatical structure, argumentation style, tone, and content organization, among other aspects. After manually reviewing the lecture note summaries, we identified five different aspects of writing preferences based on previous works (Pennebaker and King, 1999) and rhetorical structure theory (MANN and THOMPSON, 1988).

Length: The number of words in a summary.

Content Depth:

- Simple: A lecture notes overview.
- Balanced: Some detailed information on certain knowledge points from the lecture notes.
- Complex: Much detailed information from the lecture, such as formulas, code, and more.

Tongue & Style:

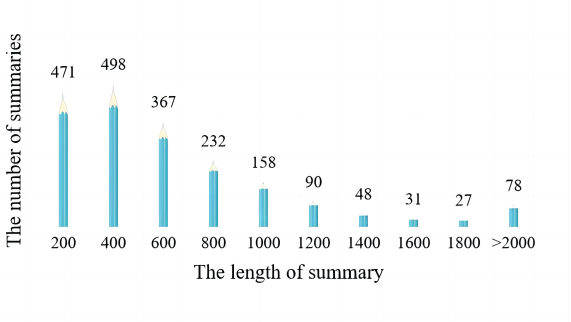

Figure 2: The length distribution of annotator-written summaries.

- Personal pronoun: Written in first-person or third-person tongue.
- Long and short sentences: Using long or short sentences in summary.

Summary Format:

- Paragraph-based: Paragraphs only in summary.
- Bullet Point: Bullet points included in paragraphs.
- Paragraph keyword: some paragraph keywords in a summary.

Rhetorical Relations:

- Elaboration: Explains, refines, or adds information to the main clause to make it clearer.
- Contrast: Compares two different viewpoints, facts, or phenomena to highlight their differences.
- Cause: Expresses a causal relationship, explaining how one event or fact leads to another.
- Background: Provides contextual information to help the reader understand the premise or background of the main clause.
- Summary: Condenses and summarizes the previous content, briefly outlining the key ideas or conclusions.

These human summary writing preferences consider not only the general linguistic features such as summary length and syntactic complexity, but also personal differences such as the individual habit for pronouns, short/long sentence use, and whether there is a structure in the summary.

Human Writing Preferences Figure 2 displays the length distribution of summaries. According to our findings, the number of summaries is the highest in the 200 to 400-word range. As the length of the summaries increases, their quantity gradually

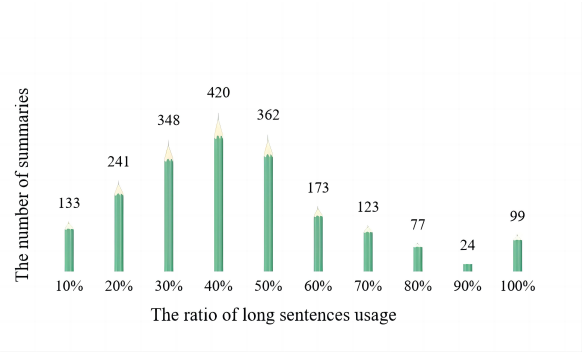

Figure 3: The ratio distribution of long sentences in annotators ' written summaries.

decreases, with shorter summaries being more common. Overall, the length of the summaries mainly concentrates between 200 and 800 words, while longer summaries are relatively rare. It reflects the *length-limitation preference*: i.e., annotators tend to use more concise summaries to convey the main information effectively.

We defined long sentences in the summaries as those exceeding 20 words and calculated the usage rate of these long sentences, as shown in Figure 3. The results indicate that only 10% of annotators did not combine long and short sentences when writing summaries. Therefore, **the combined use of long and short sentences is considered a sentence-balanced preference**.

Dimension	Preference	Percentage
Tongue&Style	First-person	0%
	Third-person	100%
Summary Format	Bullet Point	74.67%
	Paragraph-based	0%
	Paragraph Keyword	99.57%

Table 3: **When the content depth of the summary is complex,** the table displays the preferences for tongue & style in terms of personal pronoun and summary format. Notably, bullet points and paragraph keywords can coexist within the summary.

We analyze the proportions of person pronouns and summary format when composing complex summaries, as illustrated in Table 3. Notably, all annotators employed the third person to convey the objectivity of the summaries, simultaneously utilizing bullet points and paragraph keywords to enhance structural hierarchy and content detail. Consequently, the **fine-grained preference** identified is that **third-person expression is frequently paired with keywords and bullet points in drafting complex summaries.**

Dimension	Preference	Percentage
Content Depth	simple	72.64%
	balanced	25.98%
	complex	1.38%
Summary Format	Bullet Point	0.46%
	Paragraph-based	90.11%
	Paragraph Keyword	9.65%

Table 4: **When annotators use the first person to write summaries,** the table displays the proportion of different preferences in terms of content depth and summary format dimensions. Notably, bullet points and paragraph keywords can coexist within the summary.

Dimension	Preference	Percentage
Summary Format	Bullet Point	0.00%
	Paragraph-only	98.73%
	Paragraph Keyword	1.27%

Table 5: When annotators use the first person to write summaries and the content depth is simple, the table displays the proportion of different preferences in terms of the summary format dimension.

We examined the preferences for content depth and summary format when annotators utilized the first person to write summaries. Tables 4 and 5 indicate that when the first person is employed, the content of the summaries tends to be simplified, with approximately 90% of annotators relying solely on paragraphs for their composition. Further analysis reveals that when the tongue of the summary is first-person and the content is straightforward, as many as 98% of annotators restrict themselves to using only paragraphs. Therefore, we derive **the general-oriented preference** from the written summaries: **When annotators use the first person to compose summaries, the content depth is typically simple, and the summary format consists of paragraphs only.**

Rhetorical Structure	Count
Elaboration	1698
Cause	907
Contrast	893
Background	1241
Summary	1557

Table 6: Counts of Different Rhetorical Structures

Additionally, we utilize GPT-4o to analyze the

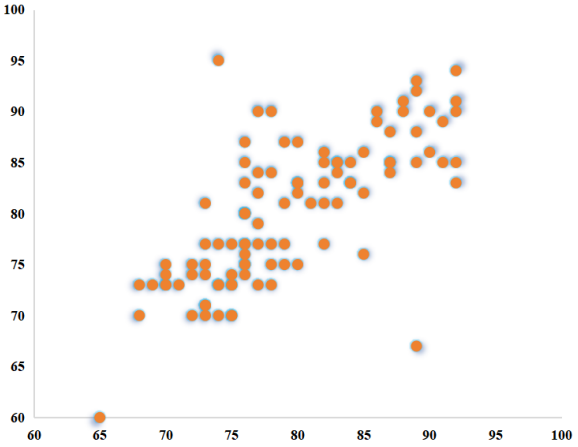

Figure 4: The relationship between two expert annotators' scores. It can seen there is a strong positive correlation between the two annotators.

rhetorical structures in human-written summaries (prompt in the appendix A), as shown in Table 6. The usage frequency of the five rhetorical structures shows a similar distribution trend, with some slight differences. **Most annotators use the "Elaboration" structure, which helps to elaborate in detail on the working principles of algorithms, models, and methods.** Due to the complexity of the machine learning field, authors typically use detailed elaboration to reinforce the understanding of core concepts, with this structure appearing 1698 times. In contrast, the "Cause" and "Contrast" structures are used less frequently, with frequencies of 907 and 893, respectively. Although these two structures play a key role in discussing model performance and comparing different algorithms, their usage scenarios are relatively limited. They typically only appear when discussing algorithms' advantages and disadvantages or comparing different methods.

3 LecSumm

3.1 Summary Quality Control

In this section, we invited two university staff who have CS PhD degrees as expert annotators to evaluate the summaries written by annotators based on the following four dimensions to ensure the quality of the data:

Coherence: The overall quality of all sentences. The summary should be well-structured and well-organized. It should not just be a collection of related information, but build coherent information about a topic from one sentence to the next.

Consistency: The factual consistency between the summary and its source. A factually consistent summary only contains statements in the source document.

Fluency: The quality of individual sentences. The sentences in the summary "should not have formatting issues, capitalization errors, or ungrammatical sentences (e.g., fragments, missing parts), which would make the text difficult to read."

Relevance: The selection of important content from the source. The summary should only include important information from the source document. Annotators are instructed to penalize summaries containing redundant and superfluous information.

We randomly selected 100 summary samples. Experts were required to score the summaries based on the above four dimensions, with a maximum of 25 points for each dimension, and calculate the total score. To assess inter-annotator agreement, we calculated Krippendorff's alpha coefficient (Gwet, 2011).

The Krippendorff's alpha score is 75.82%, indicating that the experts showed very high consistency in their annotations. Figure 4 confirms this, showing that most of the annotation scores for the summaries are between 75 and 95. These results collectively indicate that the quality of the summaries is high and that the experts exhibited a high level of consistency in their evaluations.

3.2 Dataset Construction

The data, including the provided lecture notes and human-written summaries, was extracted into plain text, removing all external information besides summaries. 200 samples were obtained after cleansing and filtering, including the ten lecture notes as fixed input, and 2,000 human-written summaries as targets. The average input document length of LecSumm is 6.5k, about one-third of the documents are over 9k, and the topics are in Table 1. We split our dataset into the train (1,600, 80%), validation (200, 10%), and test (200, 10%) subsets.

3.3 Dataset Comparison

In this section, we use four indicators to evaluate the intrinsic characteristics of datasets: coverage (Grusky et al., 2018), density (Grusky et al., 2018), redundancy (Bommasani and Cardie, 2020), and n-gram overlap. More details on the above indicators can be found in Appendix B. We chose five commonly used English long document datasets for comparison. **CNN-DM** (Nallapati et al., 2016) is a news corpus from the CNN and Daily Mail web-

Dataset	Coverage#rank	Density#rank	Redundancy#rank	% of novel n-grams			
				uni-	bi-	tri-	4-
CNN-DM	0.89#3	3.6#2	0.157#5	19.5	56.8	74.4	82.8
PubMed	0.893#4	5.6#5	0.146#4	12.4	44	65.3	76
arXiv	0.920#5	3.7#3	0.144#3	9.5	41	66.4	79.6
BigPatent	0.861#2	2.1#1	0.223#6	13.5	52.6	78.3	89.5
GovReport	0.942#6	7.7#6	0.124#2	5.7	32.7	56.3	68.9
LecSumm	0.860#1	5.5#4	0.122#1	13.3	53.3	77.4	85.6

Table 7: Intrinsic evaluations of different summarization datasets, including values and rankings, calculated on test sets only. Smaller coverage, density, and redundancy values are deemed preferable. Percentages of novel n-grams in summaries of different datasets are also provided.

sites. **PubMed and arXiv** (Nallapati et al., 2016) are from scientific papers. **BigPatent** (Sharma et al., 2019) consists of records of U.S. patent documents. **GovReport** (Huang et al., 2021) is a collection of reports published by the U.S. Government Accountability Office and Congressional Research Service.

Table 7 shows several datasets' coverage, density, redundancy, and n-gram overlap scores. Specifically, LecSumm achieves the highest scores on coverage and redundancy, meaning that less summary contents in the datasets are extracted from documents, and every summary has less repeated information, which further shows that human-centered summaries vary significantly. LecSumm's performance on the density metric is moderate due to numerous specific terms, definitions, and concepts in the lecture notes. These token sequences tend to be long and difficult to rephrase, necessitating their retention in the human-written summaries, which leads to a decrease in the density score. Nevertheless, the coverage metric indicates that LecSumm's summaries still possess the highest level of abstraction. Considering these three metrics, it is evident that LecSumm performs best in terms of abstractiveness and conciseness.

In addition to the aforementioned metrics, we further evaluate the abstractiveness of datasets. Specifically, we quantified it by calculating the percentage of novel n-grams in the summaries that didn't appear in the source text. Table 7 displays high percentages of novel tri-grams and 4-grams (Phang et al., 2023a). Combining the scores of coverage, density, and novel n-grams, it can be concluded that LecSumm possesses the best abstractiveness, making it more suitable for evaluating human-centered text summarization.

4 Experiments

We conducted a series of experiments on LecSumm to verify the usability of the dataset and answer the question: Can language models capture human writing preferences through prompt design and model finetuning?

Baselines We choose unsupervised language models: **TextRank** (Mihalcea and Tarau, 2004), **SummPip** (Zhao et al., 2020), and pre-trained language models: **LED** (Beltagy et al., 2020), **PEGASUS-X** (Phang et al., 2023b), and **LongT5** (Guo et al., 2022) as traditional models baselines. In addition, we evaluate large language models under zero-shot settings. We choose **GPT-4-turbo**, **GPT-4o** (OpenAI et al., 2024), **Qwen2.5-7B-Instruct** (Team, 2024), **Deepseek-r1-7b** (DeepSeek-AI, 2025) and **Gemma3-4b-it** (Team et al., 2025) which support 128k inputs.

Experiment settings For unsupervised language models, we used the TextRank in summanlp [2]and SummPip[3] algorithms, the two key hyperparameters in SummPip: nb_clusters and nb_words are set as 14 and 20, respectively.

For pre-trained language models, we used the led-large-16384[4], pegasus-x-large[5], and long-t5-tglobal-large[6] models for full fine tuning. We use an NVIDIA A100 80GB PCIe GPU for experiments. Models are used transformers4.35.2[7] to

[2]https://github.com/summanlp/textrank
[3]https://github.com/mingzi151/SummPip
[4]https://huggingface.co/allenai/led-large-16384
[5]https://huggingface.co/google/pegasus-x-large
[6]https://huggingface.co/google/long-t5-tglobal-large
[7]https://huggingface.co/docs/transformers/

Model	ROUGE			BertScore			UniEval			SummaC
	R-1	R-2	R-L	P	R	F	coherence	fluency	relevance	
Unsupervised model										
TextRank	12.55	5.73	5.79	77.05	80.31	78.61	69.61	75.71	67.02	**97.53**
SummPip	25.91	4.57	11.54	73.98	79.42	76.59	9.60	26.60	9.83	58.97
Fine-tuned model										
LED	15.83	4.04	10.38	80.70	79.25	79.90	49.59	76.08	50.13	68.21
PEGASUS-X	21.67	4.72	13.21	78.50	79.40	78.92	73.98	74.51	72.72	80.10
LongT5	21.67	4.74	13.22	78.50	79.41	78.91	73.98	74.51	72.73	80.02
Zero-shot LLM+Prompt(Length-limitation preference)										
GPT-4-turbo	34.13±0.91	7.51±0.41	13.98±0.41	81.41±0.17	82.39±0.09	81.89±0.09	97.51±0.33	95.12±0.41	**97.11±0.23**	50.16±10.19
GPT-4o	34.09±1.31	7.62±0.64	14.21±0.41	81.60±0.13	82.71±0.18	82.14±0.12	97.11±0.56	94.90±0.32	96.83±0.33	53.40±6.80
Qwen2.5-7b-Instruct	31.33±1.46	6.42±0.78	13.21±0.36	81.26±0.21	82.10±0.39	81.67±0.17	97.30±0.49	91.30±2.16	95.98±2.45	56.12±7.12
Deepseek-r1-7b	28.28±8.05	5.53±2.33	11.65±2.59	81.78±1.49	81.39±2.16	81.57±1.49	93.71±7.78	87.86±7.08	94.53±9.36	63.55±11.51
Gemma3-4b-it	32.92±1.31	6.70±2.13	**14.23±2.35**	81.86±1.28	81.75±1.96	81.80±1.55	95.37±5.13	94.60±1.51	94.89±2.85	58.69±11.35
Zero-shot LLM+Prompt(Sentence-balanced preference)										
GPT-4-turbo	27.88±1.94	5.58±0.47	11.97±0.60	80.73±0.21	83.12±0.29	81.90±0.16	96.69±1.09	95.25±0.28	92.96±2.53	53.21±4.80
GPT-4o	34.15±1.71	8.29±0.67	13.98±0.59	81.65±0.19	**83.22±0.21**	82.42±0.13	96.52±0.99	95.08±0.29	90.43±4.09	52.01±12.36
Qwen2.5-7b-Instruct	29.28±3.28	6.31±0.96	11.98±1.12	81.17±0.29	82.96±0.47	82.05±0.24	88.90±3.94	94.72±0.54	86.33±3.52	60.00±10.56
Deepseek-r1-7b	24.37±9.04	4.67±2.40	10.57±3.02	81.75±1.98	80.62±2.05	81.16±1.57	92.48±2.96	82.94±5.92	95.13±3.07	61.74±11.51
Gemma3-4b-it	27.91±8.14	5.36±2.16	12.65±2.95	82.55±1.27	81.24±2.08	81.88±1.55	92.48±6.18	88.08±7.98	92.71±9.13	51.79±10.47
Zero-shot LLM+Prompt(Fine-grained preference)										
GPT-4-turbo	33.55±0.83	7.59±0.32	13.43±0.18	81.44±0.13	80.76±0.28	81.09±0.17	97.67±0.69	94.14±0.84	96.74±0.62	58.52±9.25
GPT-4o	35.29±1.44	**9.10±0.76**	13.42±0.85	81.23±0.41	80.00±0.55	80.60±0.47	96.06±1.27	92.48±0.81	94.80±1.40	81.65±13.26
Qwen2.5-7b-Instruct	33.15±1.65	7.99±0.95	13.02±0.43	81.27±0.24	80.56±0.84	80.90±0.51	94.84±1.59	93.31±1.06	93.32±1.90	71.17±9.86
Deepseek-r1-7b	24.23±8.30	4.59±2.41	10.52±3.11	81.80±2.12	80.62±2.21	81.18±1.77	91.97±9.81	87.80±8.38	93.31±8.26	61.17±12.27
Gemma3-4b-it	23.13±7.63	3.70±1.81	11.14±3.21	**83.38±1.14**	80.41±1.94	81.86±1.47	86.29±1.28	95.64±5.12	89.54±10.48	53.32±9.28
Zero-shot LLM+Prompt(General-oriented preference)										
GPT-4-turbo	29.66±1.17	5.74±0.55	12.41±0.28	81.03±0.16	83.02±0.18	82.00±0.12	97.55±0.36	**95.77±0.10**	96.59±0.40	46.62±6.02
GPT-4o	31.25±2.80	7.42±0.53	13.64±2.14	81.59±0.51	83.18±0.32	82.37±0.38	97.58±0.41	95.44±0.18	90.42±7.29	58.40±8.37
Qwen2.5-7b-Instruct	32.39±3.68	6.98±1.22	13.14±1.05	81.28±0.29	82.79±0.65	82.04±0.35	97.43±0.49	93.01±0.40	95.74±0.29	62.93±12.64
Deepseek-r1-7b	26.90±7.68	5.16±2.51	10.77±2.41	79.29±1.88	81.01±1.70	80.13±1.49	94.83±2.96	82.94±5.92	95.13±3.07	66.93±9.97
Gemma3-4b-it	33.61±6.89	7.74±2.76	11.14±3.21	79.72±1.92	81.58±1.76	80.63±1.58	95.60±2.12	95.64±0.51	89.54±10.47	61.51±11.04
Zero-shot LLM+Prompt(Mixed preference)										
GPT-4-turbo	33.21±6.31	6.31±2.26	12.86±1.64	81.48±1.21	82.24±0.97	81.86±1.12	97.41±0.41	95.04±0.52	96.88±0.57	54.14±7.83
GPT-4o	**35.41±2.01**	8.55±1.14	**14.23±0.96**	82.03±0.72	83.20±0.65	**82.61±0.51**	**97.89±0.38**	95.60±0.46	96.24±0.66	64.87±10.25
Qwen2.5-7b-Instruct	30.72±2.34	6.18±1.27	13.05±1.41	81.08±1.12	82.16±0.94	81.66±0.98	96.93±1.01	92.54±1.97	95.07±1.73	59.35±8.17
Deepseek-r1-7b	26.35±6.96	4.88±2.89	10.71±2.73	80.67±1.85	80.91±2.22	80.79±1.92	92.85±6.01	86.42±6.55	93.27±6.87	60.41±11.06
Gemma3-4b-it	30.85±7.02	6.52±2.61	12.34±3.06	81.74±1.36	81.62±1.78	81.68±1.52	94.92±4.73	92.35±5.62	93.41±6.13	55.92±10.08

Table 8: It presents the evaluation results of automatic summary metrics for baseline models on the LecSumm dataset. For the zero-shot LLM experiments, due to inconsistencies in the results generated by different prompts, the standard deviation was added to the mean as a supplement.The scale of the numbers shown represents the percentages of evaluation metrics and standard deviations, with the standard deviation calculated based on the results from 10 different prompts.

finetune for ten epochs. We set the input token length as 8k and the output token length as 1024, batch_size 2, and the remaining parameters are default ones.

For zero-shot LLMs, we evaluated GPT-4-turbo[8], GPT-4o[9], Qwen2.5-7B-Instruct[10], **Deepseek-r1-7b** [11]and **Gemma3-4b-it**[12] . Given the above analysis on the four types of human preferences, we design four human writing instructions that are extracted from the annotators' summaries as instructed prompts for a zero-shot LLM. To further extend the generality of these four types of human references, we use GPT-4 to generate ten synonymous sentences for each type of human preference and mix preference. All four instruction prompt sets are in Appendix D.1. Thus, given a specific human preference, a full prompt will include a random sample of the corresponding prompt set and the input lecture note. We utilize some automatic evaluation metrics: ROUGE, BertScore (Zhang et al., 2020), UniEval (Zhong et al., 2022), and SummaC (Laban et al., 2022) to assess the summaries generated by the models and verify the usability of the dataset. More details about these metrics are in Appendix D.2.

5 Result

As shown in Table 8, TextRank and finetuned sequence-to-sequence models demonstrate superior performance when evaluated using SummaC, reflecting higher textual consistency and factual accuracy. This advantage stems from the nature of extractive and seq-to-seq models, which tend to directly copy some text spans from the input. However, these models show weaker performance on ROUGE and UniEval metrics, particularly with ROUGE scores failing to exceed 30%—a stark contrast to their performance on standard datasets like CNN/DM and GovReport (Phang et al., 2023b; Guo et al., 2022). In contrast, GPT-4 series models outperform traditional language models across most evaluation metrics (excluding SummaC), benefiting from their robust architecture and massive pretraining data. The varying performance patterns across different evaluation dimensions highlight that our constructed dataset poses challenges for both conventional summarization models and generative approaches.

Model	Length-limitation preference	Sentence-balanced preference	Fine-grained preference	General-oriented Preference
GPT-4-turbo	79.09%	96.36%	100%	100%
GPT-4o	94.55%	84.55%	100%	100%
Qwen2.5-7b-Instruct	81.82%	90%	96.36%	63.64%
Deepseek-r1-7b	60.00%	96.37%	100%	88.19%
Gemma3-4b-it	40.91%	81.82%	100%	100%

Table 9: The large language models' capture capability ratio of human writing preferences

Model	Short	Long and Short	Long
GPT-4	0.91%	96.36%	2.73%
GPT-4o	0	84.55%	15.45%
Qwen2.5-7b-Instruct	3.64%	90%	6.36%
Deepseek-r1-7b	3.63%	96.37%	0
Gemma3-4b-it	0	81.82%	18.18%

Table 10: The proportion of long and short sentences in the generated summaries by five models

Furthermore, we performed a Wilcoxon T-test to compare GPT-4-turbo and GPT-4o using each human preference prompt set separately. The resulting p-values for the four sets—length-limitation, sentence-balanced, fine-grained, and general-oriented—were 0.0613, 0.021, 0.045, and 0.043, respectively. These results indicate that there is no statistically significant difference between GPT-4-turbo and GPT-4o under the length-limitation prompt set. However, for the other three prompt sets, GPT-4o demonstrates significantly better performance than GPT-4-turbo.

6 Language models' capability on capturing human writing preferences

We further conduct human analysis of the summaries generated by the finetuned models and find that over 75% of them output summaries exceeding 800 words in length, failing to control the summary length effectively. Although the model captures the complexity of the content, it does not generate bullet points, which are present in the reference

index

[8]https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

[9]https://platform.openai.com/docs/models/gpt-4o

[10]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[11]https://huggingface.co/deepseek-ai/DeepSeek-R1

[12]https://huggingface.co/google/gemma-3-4b-it

summaries, when handling more complex content. However, for simpler summaries, the model captures the use of the first-person perspective well, with 92.45% of the generated summaries using the first-person point of view. Table 9 presents

Model	Length-limitation preference	Sentence-balanced preference	Fine-grained preference	General-oriented Preference
GPT-4-turbo	20%	100%	100%	96%
GPT-4o	28%	100%	100%	98%
Qwen2.5-7b-Instruct	69%	100%	100%	100%
Deepseek-r1-7b	41%	100%	98%	84%
Gemma3-4b-it	33%	95.4%	100%	64%

Table 11: This table shows different models capture four types of preferences: length-limitation, sentence-balanced, fine-grained, and general-oriented, under mixed-preference prompts.

the accuracy rates of five large language models in capturing four types of human writing preferences. Overall, the GPT-4 series models (GPT-4-turbo and GPT-4o) perform well across most preferences. Among them, GPT-4o stands out particularly in capturing the "length-limitation" preference, outperforming GPT-4-turbo and Qwen2.5-7 b-Instruct by more than 10%. However, in terms of the "sentence-balanced" preference (i.e., a mix of long and short sentences), GPT-4o performs relatively weakly. As shown in Table 10, GPT-4o fails to effectively respond to prompts that explicitly favor long sentences (20 words).

Notably, Deepseek-r1-7b exhibits greater performance variability. It achieves 96.37% and 100% accuracy in capturing the "sentence-balanced" and "fine-grained" preferences, respectively, but only scores 60% in "length-limitation" and 88.19% in the "general-oriented" preference. The poor performance in the length- limitation category indicates the model's difficulty in controlling the summary length. Its weaker performance on general-oriented summaries is mainly due to the frequent generation of bullet points and other structural elements, which fail to align with the natural and concise style expected in real-world summaries.

In addition, we also report the performance of five large language models in capturing four types of writing preferences under mix preference prompts, as shown in Table 11. Compared to the single-preference setting, the models' ability to capture the length-limitation preference drops significantly. For instance, the accuracy rates of GPT-4-turbo and GPT-4o fall to 20% and 28% respectively, indicating that even high-performing models struggle to balance multiple writing requirements simultaneously. In contrast, Qwen2.5-7b-Instruct maintains stable performance across all four preferences and demonstrates strong adaptability by preserving a 69% rate on length-limitation even under mixed preference prompts.

The analysis above indicates that neither fine-tuned models nor prompt-based zero-shot LLMs can effectively capture human writing preferences in real summaries. Continuous optimization and adjustment of the models are required to improve their ability to capture these preferences.

7 Conclusion

We design a lecture note summarization task to generate human-centered summaries and analyze human writing preferences across four dimensions: length, content depth, tongue & style, and summary format. We develop a novel dataset, LecSumm, which more accurately reflects specific writing preferences than publicly available datasets. Through a series of experiments involving both automatic and manual evaluations of benchmark models, we validate the effectiveness of the dataset and observe that language models still fall short of fully capturing human writing preferences, utilizing finetuning and prompt design techniques. Our findings suggest that aligning models with human preferences remains challenging, highlighting a potential direction for future research. Additionally, our dataset contributes significantly to research on human-centered text summarization.

Limitation

There are a few limitations to our work: First, the lecture notes are limited to the field of machine learning and do not cover a wide range of domains, which restricts the generalizability of our findings. Future research will expand to more diverse fields to validate the robustness and universality of the analysis. Second, due to the high cost of annotation, we were only able to recruit 200 participants. Furthermore, the use of in-context learning with large language models (LLMs) was constrained by input-output length limitations, allowing only a minimal number of examples in the experiment. This restriction hampers the model's ability to effectively learn and simulate human preferences in

summarization.

Ethics Statement

Data collection approval was received from an ethics review board. No identified personal information is collected in the data collection process. All code and data used in this paper comply with the license for use.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (Project No. 62202075, 72274096, 71774084, 72301136, and 72174087), the Foreign Cultural and Educational Expert Program of the Ministry of Science and Technology of China (G2022182009L), the Natural Science Foundation of Chongqing, China (No. CSTB2022NSCQ-MSX1404), Fundamental Research Funds for the Central Universities (No. SWU-KR24008), Key Laboratory of Data Science and Smart Education, Hainan Normal University, Ministry of Education (No. DSIE202206).

References

Sharan Narang Aakanksha Chowdhery. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

AI@Meta. 2024. Llama 3 model card.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel R. Tetreault, and Alejandro Jaimes. 2022. Mapping the design space of human-ai interaction in text summarization. *Preprint*, arXiv:2206.14863.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. *Preprint*, arXiv:1803.02324.

Kilem L Gwet. 2011. On the krippendorff's alpha coefficient. *Manuscript submitted for publication. Retrieved October*, 2(2011):2011.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Hassan Foroosh, and Fei Liu. 2023. Decipherpref: Analyzing influential factors in human preference judgments via gpt-4. Proceedings of the 2023 Conference on Empirical Methods in Natural Language ....

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

WILLIAM C. MANN and SANDRA A. THOMPSON. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisci-*

plinary Journal for the Study of Discourse, 8(3):243–281.

Danielle S McNamara, Scott A Crossley, and Philip M McCarthy. 2010. Linguistic features of writing quality. *Written communication*, 27(1):57–86.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert,

Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Tatiana Passali, Alexios Gidiotis, Efstathios Chatzikyriakidis, and Grigorios Tsoumakas. 2021. Towards human-centered summarization: A case study on financial news. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 21–27.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Jason Phang, Yao Zhao, and Peter Liu. 2023a. Investigating efficiently extending transformers for long input summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3946–3961, Singapore. Association for Computational Linguistics.

Jason Phang, Yao Zhao, and Peter Liu. 2023b. Investigating efficiently extending transformers for long input summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3946–3961, Singapore. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of gpt-4: Reliably evaluating large language models on sequence to sequence tasks. *Preprint*, arXiv:2310.13800.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun

Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jen tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *Preprint*, arXiv:2308.03656.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 12–23, Singapore. Association for Computational Linguistics.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024a. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024b. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2024c. Chatie: Zero-shot information extraction via chatting with chatgpt. *Preprint*, arXiv:2302.10205.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *Preprint*, arXiv:2309.07864.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1949–1952, New York, NY, USA. Association for Computing Machinery.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Experiment of rhetorical relations analysis

We select key rhetorical relations based on Rhetorical Structure Theory (RST), including elaboration, cause, contrast, background, and summary, and utilize GPT-4o to analyze the rhetorical structures in human-written summaries.We design the following prompt for the analysis:

Analyze the rhetorical structure of the following summary based on Rhetorical Structure Theory (RST) and identify the rhetorical relations it contains, specifically: Elaboration, Contrast, Cause, Background, and Summary.

Summary: 'summary'

Definitions of Rhetorical Relations:

- Elaboration: Explains, refines, or adds information to the main clause to make it clearer.

- Contrast: Compares two different viewpoints, facts, or phenomena to highlight their differences.

- Cause: Expresses a causal relationship, explaining how one event or fact leads to another.

- Background: Provides contextual information to help the reader understand the premise or background of the main clause.

- Summary: Condenses and summarizes the previous content, briefly outlining the key ideas or conclusions.

Please output the analysis results in the following format only, without any additional text:

Elaboration:0

Contrast: 0

Cause: 0

Background: 0

Summary: 1

B Metrics used for dataset comparison

Coverage (Grusky et al., 2018) quantifies the proportion of words in a summary that originate from an extractive fragment within the document. Its calculation method is as follows:

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

where D and S represent the document and its summary, respectively. $F(D,S)$ is the set that includes all extractive fragments. $|\bullet|$ signifies the length of a token sequence. A higher coverage score indicates that more content is directly copied from the document when generating the summary.

Density (Grusky et al., 2018) is similar to coverage, where the sum of fragment lengths is changed to the sum of squares of lengths:

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

In the event that the length of each fragment is relatively brief, the density value will be comparatively low. This implies that if two summaries share the same coverage value, the one with a lower density might exhibit greater variability because its fragments are relatively short and discontinuous.

Redundancy (Bommasani and Cardie, 2020) is used to evaluate whether sentences in a summary are similar to each other.

$$Redundancy(S) = \underset{(a,b) \in M \times M, a \neq b}{mean} R_L(x,y) \quad (3)$$

where M is sentence set of summary S, (a,b) is a sentence pair. R_L is ROUGE-L F1-score. Redundancy can be utilized to measure the degree to which sentences in a summary repeat information unnecessarily. In essence, a high-quality summary ought to strive for maximum conciseness.

C Human-written Summary Check Guidelines

Two expert annotators score summaries independently, they need to complete 100 subtasks, each of which consists of the source document and human-written summaries. We have developed a guideline for annotators, see Fig 5.

D Experiment

D.1 Prompt Design

We incorporate the human writing preferences, analyzed from real summaries, into the design of the prompts and use GPT-4 to generate 10 synonymous sentences as controls to eliminate the specificity of the prompts. These prompts correspond to four different writing preferences.

Length-limitation writing preference: Annotators tend to use more concise summaries to effectively convey the main information.

- Please generate a summary of 300-500 words in length about this lecture note.
- Please write a 300-500 word summary of this lecture note.
- Please generate a summary of 300 to 500 words based on this lecture note.

- Write a 300-500 word overview of this lecture note.
- Create a 300-500 word summary for this lecture note.
- Please summarize this lecture note in 300 to 500 words.
- Generate a 300-500 word summary about this lecture note.
- Write a summary of this lecture note that is 300-500 words long.
- Please provide a 300-500 word summary for this lecture note.
- Create a concise 300 to 500-word summary based on this lecture note.
- Summarize this lecture note in 300-500 words.

Sentence-balanced writing preference: Annotators tend to use the combination of long and short sentences.

- Please generate a summary including both long sentences and short sentences, the long sentence is over 20 words.
- Please create a summary that contains a mix of long and short sentences, with the long sentences exceeding 20 words.
- Generate a summary using both short sentences and long sentences, ensuring the long ones are more than 20 words.
- Kindly write a summary that includes both brief and extended sentences, whereas the longer ones have more than 20 words.
- Create a summary that incorporates both short and long sentences, with the longer sentences being over 20 words.
- Please write a summary that balances long and short sentences, making sure the long ones are at least 20 words.
- Provide a summary that features a combination of short sentences and longer ones, with the latter exceeding 20 words.
- Generate a summary that mixes short and long sentences, ensuring the long sentences are more than 20 words in length.
- Please draft a summary that includes both concise sentences and extended ones, with the longer sentences exceeding 20 words.
- Create a summary using both long and short sentences, ensuring the long sentences are over 20 words long.
- Please produce a summary that contains both short and longer sentences, where the long ones are more than 20 words.

Fine-grained writing preference: Third-person expression is frequently paired with keywords and bullet points in drafting complex summaries.

- Please generate a summary including more detailed information, e.g., formulas, and using third person, Bullet Point, and Paragraph Keyword.
- Please create a summary that includes more detailed information, e.g., formulas, using a third-person perspective, bullet points, and paragraph keywords.
- Kindly generate a summary with additional details, e.g., formulas presented in the third person, and incorporate bullet points and key paragraph terms.
- Please draft a summary that provides more in-depth information, e.g., formulas, using the third person, and features bullet points along with paragraph keywords.
- Create a summary with more detailed content, e.g., formulas, written in the third person and formatted with bullet points and essential paragraph keywords.
- Please write a summary including more comprehensive information, e.g., formulas, employing the third person, bullet points, and keywords for each paragraph.
- Generate a summary with greater detail, e.g., formulas, written in the third person and organized with bullet points and paragraph keywords.
- Kindly provide a summary that is more detailed, e.g., formulas, written from a third-person perspective and structured with bullet points and keywords from each paragraph.
- Please create a detailed summary, e.g., formulas, using the third person, and include bullet points as well as keywords for each paragraph.
- Draft a summary containing more detailed information, e.g., formulas, written in the third person, and incorporate both bullet points and paragraph keywords.
- Generate a summary with increased detail, e.g., formulas, using a third-person approach, and include bullet points and keywords from each paragraph.

General-oriented writing preference: When annotators use the first person to compose summaries, the content depth is typically simple, and the summary format consists of paragraphs only.

- Please generate a summary including general information and using first person, not Bullet Point and Paragraph Keywords.
- Please write a summary that includes general information in the first person without using bullet points or paragraph keywords.
- Kindly create a summary with general information, written in the first person, avoiding bullet points and keywords from paragraphs.
- Generate a summary using a first person that covers general information without incorporating bullet points or paragraph keywords.
- Please provide a summary with general details, written in the first person, and do not include bullet points or paragraph keywords.
- Draft a summary with general information in the first person, ensuring no bullet points or keywords from paragraphs are used.
- Please write a first-person summary containing general information, but exclude bullet points and paragraph keywords.
- Kindly generate a first-person summary with general content, avoiding the use of bullet points or keywords from paragraphs.
- Create a summary that includes general information, written in the first person, and does not use bullet points or paragraph keywords.
- Please produce a summary with general information using the first person, and refrain from adding bullet points or paragraph keywords.
- Write a summary in the first person that focuses on general information while not incorporating bullet points or paragraph keywords.

Mixed writing preference: We maximize the integration of four preferences and design mixed-preference prompts for implementation.

- Please produce a detailed summary of the given lecture notes, ensuring a length between 300 and 500 words. The summary should utilize a blend of both long and short sentences, with each long sentence exceeding 20 words. Incorporate comprehensive details, including necessary formulas and technical aspects. Structure the content in a third-person perspective, using bullet points to clearly outline key concepts. Additionally, introduce paragraph keywords at the beginning of each section to clarify the main discussion points.
- Generate a comprehensive summary of the provided lecture note, ranging between 300 and 500 words. The summary must contain a mix of concise and extended sentences, ensuring that longer ones have more than 20 words. It should include precise technical details, relevant equations, and concepts while maintaining a third-person narrative. Use bullet points for clarity and organize each section with paragraph keywords that indicate the core topics covered.
- Write a well-structured summary of the given lecture note, keeping its length between 300 and 500 words. The summary should be composed of both short and long sentences, with the longer ones containing at least 20 words. It must present in-depth information, including formulas and technical content. The writing should maintain a third-person perspective, employing bullet points to emphasize key points. Additionally, introduce each section with a keyword that encapsulates its primary focus.
- Create a thorough summary of the lecture notes provided, ensuring a word count between 300 and 500. The summary should balance both short and extended sentences, with each extended sentence exceeding 20 words. It must include technical specifics such as equations and formulas. The content should be written in the third-person point of view and formatted with bullet points to clearly highlight main concepts. Each section should start with a keyword that defines the main discussion area.
- Summarize the given lecture note in a structured manner, with a length between 300 and 500 words. Use a combination of both short and long sentences, ensuring that longer ones contain more than 20 words. Incorporate detailed information, including essential technical concepts and formulas. The summary should follow a third-person narrative and be formatted using bullet points for clarity. Include a keyword at the beginning of each section to indicate its central theme.
- Generate a detailed summary of the given lecture notes, ensuring the length is between 300 and 500 words. The summary should main-

tain a mix of short and long sentences, with all long sentences containing more than 20 words. It should be written in the first-person perspective and include both general information and any essential formulas or technical aspects. The text should be continuous prose without using bullet points or paragraph indicators.

- Write a thorough summary of the provided lecture material, keeping it within 300 to 500 words. The summary must have a balanced combination of short and long sentences, where every long sentence exceeds 20 words. It should incorporate general concepts as well as any relevant mathematical expressions or technical explanations, all while maintaining a first-person narrative. The structure should flow naturally without the use of bullets or explicit paragraph markers.

- Create a comprehensive summary of the lecture notes, ensuring it falls within the 300 to 500-word range. The writing should feature both concise and extended sentences, with the latter always exceeding 20 words. The summary should include overarching concepts and specific technical details, such as formulas, and should be presented in the first-person voice. The text must be continuous, avoiding the use of bullet points or paragraph indicators.

- Develop an in-depth summary of the given lecture notes with a word count ranging between 300 and 500. The text should contain a mixture of short and long sentences, with long ones always exceeding 20 words. The summary must provide both a general understanding of the topic and detailed technical aspects, including relevant formulas, while being written in the first person. The writing should be structured as seamless prose without bullet points or paragraph demarcations.

- Summarize the lecture notes in a detailed manner within 300 to 500 words, ensuring a mix of short and long sentences where the latter contain over 20 words. The summary should present both broad ideas and necessary technical details, such as formulas, all while maintaining a first-person narrative. The prose should be continuous, free from bullet points and explicit paragraph markers.

D.2 Auto evaluation metrics

We utilized some automated evaluation metrics to assess the summaries generated by the models.

Rouge We use F1-score of ROUGE-1, ROUGE-2 and ROUGE-L[13], taking into account the completeness, readability and order of summary.

BertScore (Zhang et al., 2020) computes a similarity score for each token in the candidate sentence with each token in the reference sentence. It correlates better with human judgments.

SummaC (Summary Consistency; Laban et al., 2022) is focused on evaluating factual consistency in summarization. They use NLI to detect inconsistencies by splitting the document and summary into sentences and computing the entailment probabilities on all document/summary sentence pairs, where the premise is a document sentence and the hypothesis is a summary sentence. They aggregate the NLI scores for all pairs by either taking the maximum score per summary sentence and averaging (SCZS) or by training a convolutional neural network to aggregate the scores (SCConv). We report the SCConv score and use the publicly available for implementation[14].

UniEval (Zhong et al., 2022) is a unified multi-dimensional evaluator which re-frames NLG evaluation as a Boolean Question Answering (QA) task and by guiding the model with different questions to evaluate from multiple dimensions. We report the coherence score, fluency score, and relevance score computed by UniEval[15].

E Related Work

Human-centered text summarization The human-centered text summarization approach emphasizes designing and developing summarization models that align with the needs and preferences of human users. It primarily involves human-computer interaction for building the summarization models and leverages large language models (LLMs) as evaluators to assist in assessing the quality metrics such as fluency and factual consistency of the summaries (Cheng et al., 2022; Sottana et al., 2023). Additionally, it is also applied to the construction of text summarization datasets, which involves two stages: data collection and data annotation. Existing research predominantly

[13] https://huggingface.co/docs/datasets/how_to_metrics

[14] https://github.com/tingofurro/summac

[15] https://github.com/maszhongming/UniEval

focuses on the data annotation stage, accomplished through human interaction (Gururangan et al., 2018). In contrast, human-centered data collection should prioritize simulating real-use scenarios so that the data reflects actual human needs.

Large language model for text summarization Most LLMs adopt an autoregressive structure similar to GPT, capable of automatic text summarization (ATS) (Houlsby et al., 2019). However, as the model size increased, full parameter training became costly. Research gradually shifted towards more cost-effective and efficient methods, including finetuning and prompt engineering. Prompt engineering for LLMs involves exploring and formulating strategies to maximize the use of specific functions inherent in large language models (LLMs). This process requires optimizing the input text string to more effectively leverage the LLM's intrinsic knowledge, thereby enhancing the interpretation of the input text (Liu et al., 2023). This significantly improves the quality of the generated summaries. Prompt engineering is advantageous because it does not require extensive training or rely only on a small number of samples (Narayan et al., 2021), thus reducing resource expenditure. The implementation of prompt engineering is based on methods such as template engineering, chain of thought (CoT), and agent interaction. Template engineering is another natural way to create prompts by manually creating intuitive templates based on human introspection (Zhao et al., 2023). Chain of thought (Wei et al., 2024b) is a series of intermediate reasoning steps that can significantly enhance the LLM's ability to perform complex reasoning tasks. To address issues of factual hallucinations and information redundancy in ATS, a summarization chain of thought (SumCoT) (Wang et al., 2023b) technique was proposed to guide LLMs in gradually generating summaries, helping them integrate finer-grained details from the source document into the final summary. Agents are artificial entities that perceive the environment, make decisions, and take actions (Xi et al., 2023). A three-agent generation pipeline, consisting of a generator, a lecturer, and an editor, can enhance the customization of LLM-generated summaries to better meet user expectations.

Human Written Summary Check Guidelines

This guideline is intended to give annotators a clear understanding of the task and requirements before manual annotation.Be sure to read the following content carefully.

This task is used to assess the quality of human-written summaries. You need to complete 100 tasks, each of which will provide you with an original document and a human-written summary. You need to score each summary based on four evaluation dimensions, with a maximum score of 25 points for each dimension. The four evaluation dimensions are:

- **Coherence:** The overall quality of all sentences. "The summary should be well-structured and well-organized. It should not just be a collection of related information, but should build coherent information about a topic from one sentence to the next."
- **Consistency:** The factual consistency between the summary and its source. A factually consistent summary only contains statements that are present in the source document.
- **Fluency:** The quality of individual sentences. The sentences in the summary "should not have formatting issues, capitalization errors, or obviously ungrammatical sentences (e.g., fragments, missing parts), which would make the text difficult to read."
- **Relevance:** The selection of important content from the source. The summary should only include important information from the source document. Annotators were instructed to penalize summaries containing redundant and superfluous information.

Please fill in the scores for each dimension in the table below and calculate the total score.

Task Number	Coherence	Consistency	Fluency	Relevance	Total Score

Annotation results are only used for this study. All the information will be anonymized and your personal preferences will not be disclosed.You do not have to bear any responsibility for the risk caused by your annotation results.

Figure 5: This is a human-written summary check guideline for annotators.