ALIGNSUM: Data Pyramid Hierarchical Fine-tuning for Aligning with Human Summarization Preference

Yang Han[1,2,3*] **Yiming Wang**[2*] **Rui Wang**[2†] **Lu Chen**[1,2,3] **Kai Yu**[1,2,3†]
[1]X-LANCE Lab, Department of Computer Science and Engineering, SJTU
[2] MoE Key Lab of Artificial Intelligence, SJTU AI Institute
Shanghai Jiao Tong University, Shanghai, China
[3]Suzhou Laboratory, Suzhou, China
{csyanghan,wangrui12,kai.yu}@sjtu.edu.cn

Abstract

Text summarization tasks commonly employ Pre-trained Language Models (PLMs) to fit diverse standard datasets. While these PLMs excel in automatic evaluations, they frequently underperform in human evaluations, indicating a deviation between their generated summaries and human summarization preferences. This discrepancy is likely due to the low quality of fine-tuning datasets and the limited availability of high-quality human-annotated data that reflect true human preference. To address this challenge, we introduce a novel human summarization preference alignment framework **ALIGNSUM**. This framework consists of three parts: Firstly, we construct a Data Pymarid with extractive, abstractive, and human-annotated summary data. Secondly, we conduct the Gaussian Resampling to remove summaries with extreme lengths. Finally, we implement the two-stage hierarchical fine-tuning with Data Pymarid after Gaussian Resampling. We apply ALIGNSUM to PLMs on the human-annotated *CNN/DailyMail* and *BBC XSum* datasets. Experiments show that with ALIGNSUM, PLMs like BART-Large surpass 175B GPT-3 in both automatic and human evaluations. This demonstrates that ALIGNSUM significantly enhances the alignment of language models with human summarization preferences.[1]

1 Introduction

Text summarization is a pivotal component of natural language processing, striving to produce coherent and concise summaries of textual documents (Mani and Maybury, 1999; Nenkova and McKeown, 2012; Allahyari et al., 2017). It can be categorized into two main styles: *Extractive summarization* (Nallapati et al., 2017; Zhou et al., 2020; Zhong et al., 2020) involves selecting significant portions of the text directly from the source; In contrast, *abstractive summarization* (See et al., 2017; Lewis et al., 2019) involves generating new text that conveys the original content's essential meaning. Studies in this field often train Pre-trained Language Models (PLMs) (Vaswani et al., 2017; Radford et al., 2018; Lewis et al., 2019; Raffel et al., 2020) on standard datasets such as *CNN/DailyMail* (Nallapati et al., 2016) and *BBC XSum* (Narayan et al., 2018) to fit summary features. They usually report the performance with reference-based *automatic scores* such as ROUGE (Lin, 2004), which directly compare generated summaries with gold summaries, and fine-grained *human ratings*, which actually reflect underlying human preferences.

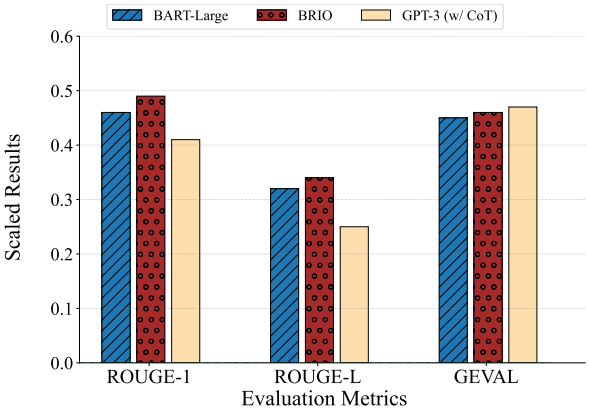

Figure 1: Results (scaled to 0-1) of automatic score ROUGE (Lin, 2004) and human rating GEval[2](Liu et al., 2023a) on the standard dataset *CNN/DailyMail*. It is obvious that PLMs perform better than LLMs on automatic scores but worse on human ratings.

However, recent investigations (Goyal et al., 2022; Wang et al., 2023c) have revealed inconsistencies between automatic scores and human ratings for both PLMs and large language models (LLMs). As shown in Figure 1, when com-

*Yang Han and Yiming Wang contribute equally.
†Rui Wang and Kai Yu are the corresponding authors.
[1]The code is released at: https://github.com/csyanghan/AlignSum

[2]Four aspects: Coherence (score range: 1-5), Consistency (score range: 1-5), Fluency (score range: 1-3), Relevance (score range: 1-5). We use GPT-4 rating that closely aligns with human judgments (Liu et al., 2023a; Wang et al., 2023a).

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 8506–8522
November 12-16, 2024

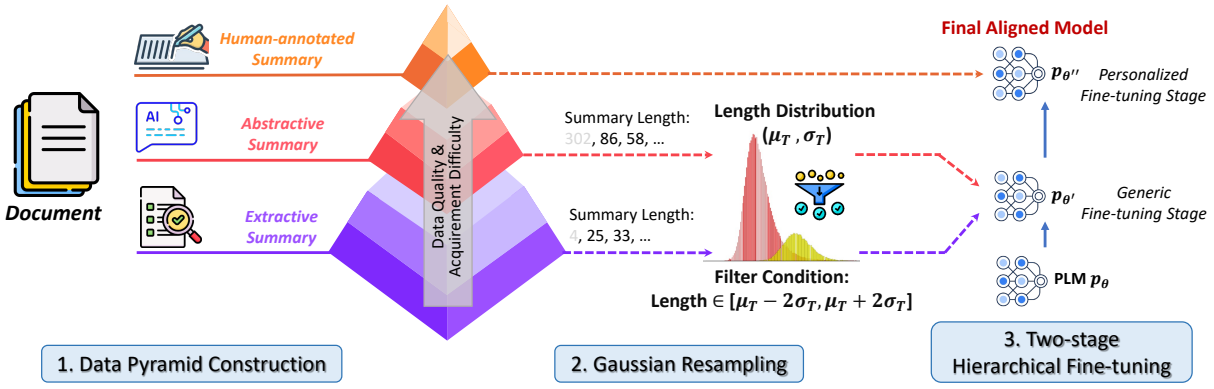

Figure 2: The overall pipeline of our summarization preference alignment framework ALIGNSUM.

pared to LLMs like GPT-3 (with Chain-of-Thought, CoT) (Wang et al., 2023c), PLMs like BART-Large (Lewis et al., 2019) and BRIO (Liu et al., 2022) fine-tuned on the *CNN/DailyMail* demonstrate impressive performances on automatic scores exceeding LLMs, but poor performances on human ratings. This contradiction stems from that PLMs are fitting low-quality summary data (Wang et al., 2023c), indicating that they need more high-quality data for aligning with human preferences to perform better in human ratings.

On the other hand, annotating a large number of high-quality summary datasets is impractical: (1) Regarding the time cost, the average reading rate for a native English speaker is approximately 220 words per minute (Gleni et al., 2019; Brysbaert, 2019). Moreover, summarization involves a structured cognitive process: reading, comprehending, and summarizing, annotators often spend twenty minutes or more to write a single summary (Wang et al., 2023c), rendering the annotation process time-consuming; (2) Regarding the labor cost, ensuring the accuracy and consistency of summaries requires cross-verification, which demands considerable human and financial resources (Ahuja et al., 2021; Zhang et al., 2023d; Chen et al., 2023). These potential obstacles collectively contribute to the scarcity of high-quality summary data.

From this consideration, instead of traditional naive fine-tuning on large amounts of training data, we would like to fully use the extremely limited amount of high-quality data to push the upper limit of PLMs' summarization ability. To address this problem, we propose **ALIGNSUM**, a novel summarization preference alignment framework. First, we design a bottom-to-up data construction method **Data Pyramid (DP)**, which consists of three components: extractive data, abstractive data, and human-annotated data. Different levels of data are collected with different methods. DP is the core component of the alignment framework, after obtaining DP, we design the **Gaussian Resampling** technique to smooth the length distribution of all summaries, and the two-stage **Hierarchical Fine-Tuning (HFT)** to maximize the use of low-resource high-entropy human preference summary data.

We conduct experiments on human-annotated *CNN/DailyMail* and *BBC XSum* datasets proposed by Wang et al. (2023c), which reflects the implicit element-aware human writing preference. We find that the pre-trained BART-Large applied with AlignSum surpasses 175B GPT-3 on both automatic scores and human ratings, achieving amazing results in outperforming large models with small models and small amounts of preference data.

2 ALIGNSUM: Summarization Preference Alignment Framework

We first formalize the summarization task: Given a document $D = d_1 d_2 ... d_n$ with length n, the goal is to generate a summary $S = s_1 s_2 ... s_m$ with length m, and usually $m \ll n$. Our proposed preference alignment framework consists of three parts: Data Pyramid (DP), Gaussian Resampling, and Two-stage Hierarchical Fine-tuning (HFT).

Figure 2 shows the overall framework: Firstly, we construct the Data Pyramid using various methods such as extraction, LLM generation, and human annotation. Secondly, as the source data have different summary lengths, PLMs with this data would lead to inconsistent summary lengths. To address this issue, we utilize Gaussian Resampling to adjust the generated summary lengths to approximate the target length. Finally, we apply

Style / Type	Difficulty	Volume
Extractive	Easy	Large
Abstractive	Medium	Small
Human-annotated	Hard	Little

Table 1: Features of summaries in Data Pyramid: summary style / type, acquisition difficulty, and data volume.

a two-stage hierarchical fine-tuning strategy: initially training the PLMs on extractive and abstractive data to fit the general domain, followed by fine-tuning the justly fine-tuned PLMs on human-annotated data to align with human preference. Details will be introduced in the following parts.

2.1 Data Pyramid Construction

Data Pyramid comprises three levels: extractive, abstractive, and human-annotated data. From bottom to top, they are arranged in increasing quality and access difficulty, while the quantity decreases (as shown in Table 1). The first two are the two most generic styles in the summarization field, and we refer to them collectively as *generic data*; the last is the most critical part used to align human preferences, and we refer to it as *personalized data*.

Extractive Data. The extractive data constitutes the majority of the pre-training corpus and is the easiest to acquire. We adopt the GSG technique proposed by Zhang et al. (2020) to select the most important sentence as the pseudo summary $\hat{S}$:

$$\begin{aligned} r_i &= \text{Rouge}(d_i, D_{\setminus d_i}), \\ \hat{S} &= \text{argmax}_{d_i}\{r_i\}_{i=1}^{n}. \end{aligned} \tag{1}$$

We use the ROUGE-1 metric (Lin, 2004) to calculate the similarity and iterate through the entire document to find the most similar sentence as the pseudo summary. Unlike the method described by Zhang et al. (2020), we extract only a single sentence due to the variability in sentence lengths, as controlling by the number of sentences is unreliable. Instead, sample selection is based on the number of tokens in the Gaussian resampling stage.

Abstractive Data. The extractive data helps identify important sentences within a document but is insufficient for summarizing crucial information that spans multiple sentences. In contrast, LLMs are effective zero-shot summarizers, capable of extracting summary information across sentences and at the document level (Goyal et al., 2022; Zhang et al., 2023c). We use both system and user prompts to guide LLMs in summarizing the document D and and generating the pseudo summary $\hat{S}$. As shown in Table 2, the system prompt specifies general requirements for accurate summarization. The document is then inserted before the user prompts, ensuring the LLM can read the entire document and adhere to user requirements. The user prompt is dataset-specific, setting the desired summary length and number of words.

Document	*E.g.*: Newcastle stand-in skipper Moussa Sissoko is facing disciplinary action after he was sent off following a reckless challenge on Liverpool midfielder Lucas ...
System Prompt	Generate a concise and coherent summary towards the given article and don't generate anything else. Make sure the summary is clear, informative, and well-structured.
Dataset-specific User Prompt	Summarize the article in `[sent num]` sentences around `[word num]` words.

Table 2: Zero-shot Summarization prompt to generate Abstractive Data with LLM.

Human-annotated Data. Human-annotated data is the most critical component of DP for aligning with human preference. Training on data generated by adapted GSG and LLMs has allowed PLMs to acquire domain-specific knowledge. However, to generate summaries that align with human preferences, further fine-tuning on annotated data is necessary. This annotated data contains explicit user preferences and is easy to acquire without specific instructions, as PLMs can learn preferences through the data itself. To avoid the variability of random annotations, we use the Element-aware dataset provided by Wang et al. (2023c). This dataset adheres to specific instructions, incorporating both micro and macro demands (Details refer to Appendix B.1), ensuring consistent and high-quality human annotations.

2.2 Gaussian Resampling

DP draws from three distinct data sources, each with unique token length distributions for their pseudo summaries. As shown in Figure 3, there are noticeable differences in summary token length distributions of extractive and abstraction data. Therefore, training directly with these disparate distributions can result in overly long or short summaries.

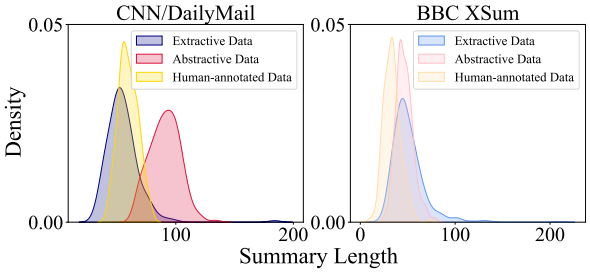

Figure 3: Summary token length distributions of DP.

To address this issue, we introduce the Gaussian Resampling technique to align all summary lengths with human-annotated summaries. Specifically, we model the token length distribution of human-annotated data as a Gaussian distribution:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{2}$$

where μ and σ represent the mean and standard deviation, respectively. With a 95% probability, the confidence interval for the token length distribution is $[\mu - 2\sigma, \mu + 2\sigma]$. We resample extractive and abstractive data within this interval to remove samples with excessively long or short pseudo summaries.

2.3 Two-stage Hierarchical Fine-tuning

Now we have obtained the resampled DP, a naive strategy is to fine-tune PLMs with them to enhance their summarization ability and align them with human preference simultaneously. However, this process can be challenging because the small amount of high-entropy data, which is crucial for alignment, can be interfered with by information from a large amount of low-entropy data (Wang et al., 2023b), leading to the underutilization of DP.

To avoid this potential issue, we propose a two-stage hierarchical fine-tuning strategy. Give a PLM p_θ, First is the **generic fine-tuning stage**, where we fine-tune p_θ with the extractive and abstractive data to enhance its ability to generate domain-general summaries, obtaining a model $p_{\theta'}$. Next is the **personalized fine-tuning stage**, where we fine-tune $p_{\theta'}$ with the human-annotated data to create the final model $p_{\theta''}$ aligned with human preferences.

Why Hierarchical Fine-tuning? From a theoretical perspective of uncertainty reduction, we can explain the advantages of hierarchical fine-tuning using DP over hybrid fine-tuning. We denote X, Y, Z as the pre-trained data (intrinsic data of PLMs), generic data (extractive/abstractive data), and personalized data (human-annotated data), respectively. $p_{x;\theta}, p_{x,y;\theta}, p_{x,y,z;\theta}$ are models after pre-training, generic fine-tuning, and personalized fine-tuning, respectively. Let $J(p_\theta)$ denote a random variable reflecting the summarization preference alignment ability of p_θ, it is obviously that

$$\begin{aligned} J(p_\theta = p_{x;\theta}) &< J(p_\theta = p_{x,y;\theta}) \\ &< J(p_\theta = p_{x,y,z;\theta}). \end{aligned} \tag{3}$$

In general, generic data enhances the performance of downstream tasks, whereas task-specific data compromises the generalized capabilities of the model, *i.e.*, the "Alignment Tax" (Ouyang et al., 2022; Dong et al., 2023). Therefore, we can intuitively make the following assumptions about the relationship between alignment uncertainty and alignment ability $J(p_\theta)$ of model p_θ:

Assumption 2.1 *For hierarchical data $\{X, Y, Z\}$, data at the lower level of DP enhances the model's ability of the upper-level tasks, but data at the upper level impairs the model's ability of the lower-level tasks, i.e.,*

$$\begin{aligned} &1.\ H(Z|J(p_\theta = p_{x;\theta})) > H(Z|J(p_\theta = p_{x,y;\theta})) \\ &\qquad > H(Z|J(p_\theta = p_{x,y,z;\theta})), \\ &2.\ H(Y|J(p_\theta = p_{x;\theta})) > H(Y|J(p_\theta = p_{x,y;\theta})), \\ &3.\ H(Y|J(p_\theta = p_{x,y;\theta})) < H(Y|J(p_\theta = p_{x,y,z;\theta})) \\ &4.\ H(X|J(p_\theta = p_{x;\theta})) < H(X|J(p_\theta = p_{x,y;\theta})) \\ &\qquad < H(X|J(p_\theta = p_{x,y,z;\theta})). \end{aligned} \tag{4}$$

We derive the uncertainty reductions before and after fine-tuning for both fine-tuning strategies:

- hybrid fine-tuning:

$$\begin{aligned} G_{hy} = &|H(Y, Z|J(p_\theta = p_{x,y,z;\theta})) \\ &- H(Y, Z|J(p_\theta = p_{x;\theta}))| \end{aligned} \tag{5}$$

- hierarchical fine-tuning:

$$\begin{aligned} G_{hi} = &|\underbrace{H(Y|J(p_\theta = p_{x,y;\theta})) - H(Y|J(p_\theta = p_{x;\theta}))}_{\text{generic fine-tuning stage}}| \\ &+|\underbrace{H(Z|J(p_\theta = p_{x,y,z;\theta})) - H(Z|J(p_\theta = p_{x,y;\theta}))}_{\text{personalized fine-tuning stage}}| \end{aligned} \tag{6}$$

We can prove that $G_{hi} > G_{hy}$ holds constant for any model p_θ and data sets X, Y, Z under the Assumption 2.1. This means the uncertainty reduction from hierarchical fine-tuning is greater, leading to a better alignment performance. Appendix A shows the complete proof. Table 6 in Section 4.1 also demonstrates the need for hierarchical fine-tuning from an empirical perspective.

Model \ Dataset	*CNN/DailyMail*				*BBC XSum*			
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Direct Generation (w/ LLMs)								
175B GPT-3, 0-shot	*42.98*	*19.48*	*28.33*	*0.8943*	*38.50*	*15.09*	*29.09*	*0.8981*
w/SumCoT, 0-shot	*49.73*	*26.10*	*36.29*	*0.9080*	*44.36*	*19.93*	*34.70*	*0.9053*
GPT-3.5-Turbo, 0-shot	*41.82*	*18.50*	*27.63*	*0.8958*	*31.38*	*13.37*	*23.05*	*0.8865*
w/Style, 0-shot	*45.62*	*19.51*	*31.52*	*0.8997*	*41.80*	*18.31*	*31.58*	*0.8984*
w/Style, 1-shot	*45.71*	*18.70*	*29.98*	*0.8996*	*41.32*	*17.19*	*31.52*	*0.8985*
LLaMA-2-7B	*44.78*	*18.83*	*29.65*	*0.8985*	*37.99*	*14.20*	*28.72*	*0.8952*
LLaMA-3-8B	*46.27*	*20.23*	*31.23*	*0.9011*	*40.34*	*16.12*	*30.00*	*0.8959*
Naive Fine-tuning (w/ PLMs)								
BART-Base	44.67	20.43	29.86	0.8754	30.04	8.95	21.71	0.8787
BART-Large	46.01	21.92	32.08	0.8851	28.73	8.80	20.96	0.8811
T5-Large	43.64	19.23	30.76	0.8842	29.83	9.14	21.99	0.8790
PEGASUS	41.39	15.66	27.26	0.8706	29.26	7.56	21.26	0.8825
BRIO	46.66	22.35	31.01	0.8876	28.45	8.34	21.05	0.8787
ALIGNSUM (w/ PLMs, Ours)								
LLaMA-2-7B (w/ HD)	44.37	18.17	28.96	0.8906	37.08	14.07	28.57	0.8937
BART-Large (w/ HD)	46.57	21.97	32.00	0.9040	40.19	14.95	28.74	0.8915
BART-Base (w/ full DP)	45.01	20.51	31.79	0.8998	39.88	16.46	30.45	0.8911
BART-Large (w/ full DP)	**48.83**	**24.11**	**34.16**	**0.9058**	**42.38**	**17.75**	**31.64**	**0.8962**

Table 3: Automatic metrics ROUGE-1/2/L and BERTScore Performances of LLMs and PLMs under naive fine-tuning and our ALIGNSUM settings on human preference Element-Aware dataset (Wang et al., 2023c). *Italic* means LLM results inferred via API or pre-trained weights, details are shown in Appendix B.2. **Bold** represents the best performances among all fine-tuned models, "w/ style" means style control with prompt in Table 2, "w/ HD" indicates fine-tuning with HD data, and "w/ full DP" represents our final model. The result of BART (w/ HD) is sampled 5 times and reports the mean. Details are shown in Appendix D.1.

3 Experiments

3.1 Setup

Dataset. We conduct DP construction and experiments on two extensively used news datasets, *CNN/DailyMail* (Nallapati et al., 2016) and *BBC XSum* (Narayan et al., 2018). For generation of extractive data (ED) and abstractive data (AD), we divide the standard training set with an 8:2 ratio to generate ED and AD for training, respectively. For human-annotated data (HD) that implicitly reflect human preference[3], we adopt the Element-Aware *CNN/DailyMail* and *BBC XSum*, which is the high-quality rewritten version (Wang et al., 2023c) of the two datasets (each 200 samples). Refer to Appendix B.1 for detailed preference features, and Appendix B.3 for data examples of the two datasets. For testing, we randomly split HD into a training set and a test set, each containing 100 samples.

Data Statistics. Table 4 shows the total count and token length distribution of pseudo summary in DP. The training set and test set are randomly sampled from the Element-Aware dataset. ED extracts the most important sentence from the original document, and the token length varies greatly. After the Gaussian Resampling, the ED_r standard deviation slows down. Although the mean length of ED_r is smaller than the HD, it all falls into the HD's distribution confidence interval. The same for AD_r, standard deviation slows down and all data token lengths fall into the desired range.

Baselines. We choose two settings for baselines: (i) Zero-Shot Generation with LLMs, we select 175B GPT-3 (Brown et al., 2020) and GPT-3.5-Turbo; (ii) Naive Fine-tuning with PLMs, means directly fine-tuning models with standard training sets of corresponding datasets. We select BART-Large, BART-Base (Lewis et al., 2019), T5-Large (Raffel et al., 2020), PEGASUS (Zhang et al., 2020), BRIO (Liu et al., 2022), LLaMA-2-7B (Touvron et al., 2023), and LLaMA-3-8B (MetaAI, 2024). All model weights are downloaded from HuggingFace. Refer to Appendix B.4 for more details.

Implementation. We use the pre-trained BART-Large for the backbone of ALIGNSUM and LLaMA-2-7B for generating abstractive data due to its ease of use, with Appendix B.5 showing that different LLMs perform similarly. For PLMs, we truncate

[3]These preferences reflect professional implicit writing styles embedded in texts and are difficult to capture explicitly.

Data	*CNN/DailyMail*	*BBC XSum*
	Sample Number / Length Mean$_{\pm std}$	
Training Set		
ED	229k / 55$_{\pm 17}$	163k / 51$_{\pm 53}$
ED$_r$	224k / 54$_{\pm 12}$	107k / 41$_{\pm 7}$
AD	57k / 91$_{\pm 13}$	41k / 46$_{\pm 9}$
AD$_r$	40k / 85$_{\pm 8}$	32k / 43$_{\pm 6}$
HD	0.1k / 64$_{\pm 17}$	0.1k / 34$_{\pm 10}$
Test Set		
HD	0.1k / 66$_{\pm 15}$	0.1k / 33$_{\pm 8}$

Table 4: Sample numbers and pseudo summary token length statistics. We use BART-Large as the tokenizer. ED$_r$ and AD$_r$ mean ED and AD after Gaussian Resampling, respectively. Data colored by gray are not involved in the actual training process.

documents to 1024 tokens and target summaries to 128 tokens following Zhang et al. (2020). Given that LLMs can handle up to 4096 tokens, we truncate the original documents to 2048 tokens for LLM inference. To ensure a fair comparison, we fine-tune all PLMs with both extractive and abstractive data for 3 epochs, using a learning rate of $5e^{-5}$ and a batch size of 128. Due to the limited amount (only 100 samples) of human-annotated data, we fine-tune them with 20 epochs, keeping the other hyperparameters unchanged.

3.2 Automatic Evaluation

Automatic evaluation usually contradicts human evaluation when referenced gold summaries are low-quality (Goyal et al., 2022). However, when references are high-quality, automatic evaluation results are more consistent with that of human evaluation as verified by Wang et al. (2023c).

Table 3 presents the overall results:

Comparisons with Naive Fine-tuned PLMs. Compared with SOTA results of PLMs under the naive fine-tuning setting, BART-Large with ALIGNSUM improves ROUGE-1/2/L by over +2.17/+1.76/+2.08 points on *CNN/DailyMail* and by +12.37/+8.61/+9.65 points on *BBC XSum*, even though these models are pre-trained with the original low-quality dataset. BERTScore for BART-Large (w/ full DP) is also higher than for the other PLMs. These results indicate that: (1) Fine-tuning on low-quality original datasets does not enhance human alignment; (2) Further fine-tuning on HD data significantly boosts performance, as seen with BART-Large (w/ HD) improving BART-Large (w/ Naive Fine-tuning) by nearly +0.5 points and +12 points on *CNN/DailyMail* and *BBC XSum*.

Comparisons with Zero-shot LLMs. Compared to LLMs with the zero-shot setting, since summarization is unsuitable for few-shot due to restricted context, we find that even though part of DP is generated from LLAMA-2-7B, its ROUGE and BERTScore are lower than BART-Large (w/ full DP). Additionally, GPT-3 performs worse than LLAMA-2-7B because we control the generation length in Section 2.1, whereas GPT-3 is only prompted with "*Summarize the above article*" as used in (Goyal et al., 2022; Sanh et al., 2021). BART-Large (w/ full DP) is slightly worse than GPT-3 (w/CoT), which is expected since GPT-3 was carefully prompted according to the data annotation protocol, making it less adaptable to other writing styles. In contrast, our model aligns with specific human preferences using modest HD data.

Comparisons with Fine-tuned LLMs. Compared to fine-tuning LLMs, we use LoRA (Hu et al., 2021) to fine-tune LLaMA2-7B. Despite LoRA having significantly fewer trainable parameters than BART fine-tuning, its memory consumption during training exceeds that of BART, even with a batch size of 1. This makes training unfeasible on consumer-grade hardware. Additionally, fine-tuning with only 100 HD samples fails to improve performance and may even decrease it, as shown in Table 3. This is because high-quality fine-tuning typically requires datasets on the order of tens of thousands (Deng et al., 2023; Zhao et al., 2024). Furthermore, this fine-tuning process may negatively impact the LLMs' other capabilities, such as mathematical and logical reasoning.

Significance Test. Given the limited sample size of 100 for each dataset, we apply Analysis of Variance (ANOVA, St et al. (1989)) to determine if statistically significant differences exist between the random experiments. Table 5 indicates significant differences, as nearly all p-values are below 0.05 ($p < 0.05$).

Dataset	ROUGE-1	ROUGE-2	ROUGE-L
CNN/DailyMail	0.013	2.59e-6	5.89e-4
BBC XSum	0.022	0.056	0.028

Table 5: ROUGE-1/2/L p-values of multiple experiments on *CNN/DailyMail* and *BBC XSum*.

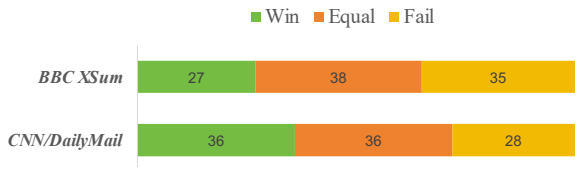

Figure 4: Reference-based human evaluation of BART (w/ full DP) and GPT-3 (w/CoT) compared to the golden reference on *CNN/DailyMail* and *BBC XSum*.

3.3 Human Evaluation

We conduct human evaluations to compare the performances of PLMs with ALIGNSUM and 175B GPT-3 (w/CoT) for it is the strongest LLM in automatic evaluation. Typically, human evaluation is reference-free [4] and involves in informativeness, conciseness, readability, and faithfulness (Bao et al., 2023; Liu et al., 2023a). We instead use a reference-based evaluation for two reasons: (1) The Element-Aware dataset has included expert-written high-quality references; (2) Referenced summaries represent a specific writing style, and evaluating only the four qualities would overlook implicit preference features captured by HD.

Given generated summaries of BART-Large with ALIGNSUM and 175B GPT-3 (w/CoT), and expert-written high-quality reference summaries, human evaluation follows these instructions:

- *Length Pre-screening*: Summaries that are too long or short compared to the reference text are considered "Fail". If both generated summaries "Fail", they are considered "Equal".

- *Overall Evaluation*: If generated summaries have similar lengths, we compare their informativeness. Informativeness is defined by characteristic elements: entities, dates, events, and results (Wang et al., 2023c), each denoted as a set (S_{en}, S_{da}, S_{ev}, S_{re}) for a summary S. Let generated summaries of ALIGNSUM and 175B GPT-3 (w/CoT) as $\hat{S}_1$ and $\hat{S}_2$ and the gold reference as G, we can define informativeness for each generated summary:

$$\text{Info}_i = \sum_j |G_j \cap [\hat{S}_i]_j|, \ i = 1, 2 \\ j \in \{\text{en}, \text{da}, \text{ev}, \text{re}\}, \tag{7}$$

  where $|\cdot|$ represents the number of elements in the set. If $\text{Info}_1 > \text{Info}_2$, then ALIGNSUM "Win"; if $\text{Info}_1 = \text{Info}_2$, they are "Equal"; otherwise, ALIGNSUM "Fail".

[4] To evaluate ALIGNSUM more comprehensively, we also conduct a reference-free human evaluation, with details shown in Appendix C.

Component			Metric		
DP	GR	HFT	ROUGE-1	ROUGE-2	ROUGE-L
CNN/DailyMail					
			41.38	18.35	26.05
✓			39.01	14.59	25.83
✓	✓		37.63	13.86	24.75
✓		✓	47.53	21.78	32.56
✓	✓	✓	**48.39**	**23.33**	**34.47**
BBC XSum					
			34.86	12.22	24.19
✓			38.46	16.79	28.68
✓	✓		39.71	16.92	28.51
✓		✓	**44.58**	19.60	**32.90**
✓	✓	✓	43.68	**19.73**	32.15

Table 6: Ablation study on the effectiveness enhancement from different components of ALIGNSUM, including Data Pyramid (DP), Gaussian Resampling (GR), and Hierarchical Fine-Tuning (HFT).

We recruited one Ph.D. student and two Master students majoring in Computer Science to conduct the evaluation following the above instructions. The majority vote is selected as the final rating, and the human evaluation results are presented in Figure 4. Although BART-Large with ALIGNSUM slightly underperforms in automatic evaluation compared to GPT-3 (w/CoT), it achieves "Win" and "Equal" ratings of up to 65% and 72% on the *BBC XSum* and *CNN/DailyMail* datasets, respectively. This demonstrates that **our ALIGNSUM can effectively align PLMs with human evaluation standards without requiring billions of model parameters or sophisticated prompt designs**. Additional case studies are provided in the Appendix D.2.

4 Ablation Study

4.1 Components of ALIGNSUM

We conduct ablation for ALIGNSUM's components to verify their effectiveness. Table 6 shows the results under different component combinations.

Gaussian Resampling. When adding the Gaussian resampling component, the performance on *CNN/DailyMail* improves, where "DP+GR+HFT" improves upon "DP+HFT" by +0.86/+1.55/+1.91 points in ROUGE-1/2/L, respectively. However, when on *BBC XSum*, we observe a slight performance degradation. This may be attributed to the

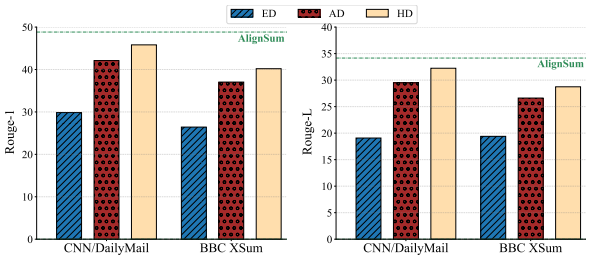

Figure 5: ROUGE-1/L of fine-tuning BART-Large with ED, AD, HD on *CNN/DailyMail* and *BBC XSum*.

HD size	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
10	44.72	18.96	29.48	0.8855
50	47.38	21.67	31.73	0.8897
100	48.04	22.67	33.38	0.9050

Table 7: ROUGE-1/2/L and BERTScore results on *CNN/DailyMail* under various HD sizes.

raw data distribution closely matching the target distribution, while the Gaussian Resampling filters out nearly 20-30% of the raw data.

Hierarchical Fine-Tuning. When adding the Hierarchical Fine-Tuning component, we observe substantial improvements in both datasets, with an average of +10 points improvement on *CNN/DailyMail* and +4 points improvement on *BBC XSum*. This validates the conclusion proved in Section 2.3 that the two-stage fine-tuning helps to reduce the information loss of high-entropy variables and maximize the use of the limited preference summary data, and also demonstrates that mixed fine-tuning tends to dilute the impact of HD within the larger volumes of ED and AD.

4.2 Components of Data Pyramid

We fine-tune BART-Large with ED, AD, and HD separately. Figure 5 shows the ROUGE-1/L results on the two datasets. The importance of high-quality data becomes increasingly evident, as fine-tuning with any single data type cannot outperform our proposed framework with DP.

4.3 Human-annotated Data (HD) Size

Table 7 illustrates the impact of varying amounts of human-annotated data on BART's ability to learn user summary patterns. With 50 training samples, BART's performance already surpasses that of all pre-trained models and LLMs. Furthermore, as the amount of human-annotated data increases, the model's performance improves correspondingly.

Fixed Sample Pool			
τ	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
0.2:0.8	43.69	19.73	32.15
0.5:0.5	44.20	**21.14**	**33.88**
0.8:0.2	43.00	19.43	33.11
1:0	44.13	19.50	32.93
Increasing Sample Pool			
σ	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
0.2	44.38	19.89	33.43
0.5	43.10	19.48	32.46
0.8	**44.56**	20.20	33.29
1	43.56	18.84	32.64

Table 8: Results of scaling abstractive data (AD) on *BBC XSum*. ***"Fixed Sample Pool"*** refers to maintaining the original training set, where τ denotes the proportion used to generate AD and ED. In contrast, ***"Increasing Sample Pool"*** indicates that the original training data is utilized twice: initially for generating ED, with σ representing the proportion used to generate AD.

4.4 Abstractive Data (AD) Scaling

High-quality HD is difficult to acquire, but AD is also useful and relatively easier to generate. Table 8 presents the results of scaling AD on *BBC XSum*. When keeping the training sample size fixed while varying the proportion of AD and ED, it is evident that increasing the amount of AD does not enhance performance. In fact, ED plays a critical role in improving model performance, as performance significantly degrades when $\tau = 1:0$, which indicates exclusive use of AD. In this setting, increasing AD would lead to a reduction in ED, introducing multiple variable changes. Then, we fix the ED component and utilize the entire training data to generate them, while we use different proporation of training data to generate AD again, which would increase the total training sample size. However, the results again demonstrate that increasing AD does not improve performance, due to a mismatch between the data distribution of AD generated by large LLMs and the test set. In contrst, DP enhances performance by introducing greater diversity, as ED effectively identifies key sentences, while AD struggles to do so.

5 Related Work

Extractive Summarization. Extractive summarization aims to extract sentences from given documents (Zhong et al., 2020). Current approaches for-

mulate this task as a classification or matching problem using recurrent neural networks (Cheng and Lapata, 2016; Nallapati et al., 2016), pre-trained language models (Liu and Lapata, 2019; Wang et al., 2022), large language models (Zhang et al., 2023b), and even diffusion models (Zhang et al., 2023a). Although extractive summarization cannot effectively synthesize summary information across sentences at the document level, they are always grammatically correct and faithful to the original text. Therefore, we utilize extractive summarization as the basis to help identify key sentences in the original document and generate extractive data.

Abstractive Summarization. Abstractive summarization generates summaries using novel phrasing and sentence fusion or paraphrasing techniques (Shen et al., 2023; Xiao et al., 2022). The seq2seq framework (Sutskever et al., 2014) with encoder-decoder architectures based on RNNs (Chung et al., 2014; Hochreiter and Schmidhuber, 1997) and Transformers (Vaswani et al., 2017) are dominant in this field. Recently, there has been a surge in prompting LLMs such as GPT (Brown et al., 2020). Studies like Goyal et al. (2022) have investigated the performance of GPT-3 and fine-tuned models, finding that the former is more preferred by humans despite having lower ROUGE scores. Zhang et al. (2023c) iteratively refines summaries through self-evaluation and feedback, exploring the use of knowledge and topic extractors to enhance summary faithfulness and controllability. Liu et al. (2023b) finds that LLMs generate summaries preferred by humans and proposes improving PLMs using LLMs as references through supervised fine-tuning and contrastive learning. In this paper, we also utilize the zero-shot summarization capability of LLMs to comprehensively understand entire documents and generate abstractive data. However, we find that LLM-generated summaries alone are not optimal, and incorporating more diverse data better aligns with human preferences.

Domain Adaptation Summarization Domain adaptation summarization has been widely studied in low-resource settings (Yu et al., 2021; Balde et al., 2024; Fabbri et al., 2020). Gururangan et al. (2020) demonstrates that domain- and task-adaptive pretraining consistently improves performance, though their work focuses on eight classification tasks and relies solely on pretraining with unlabeled data. WikiTransfer (Fabbri et al., 2020) extends domain adaptation to summarization by fine-tuning pretrained models on pseudo-summaries generated from Wikipedia data using ROUGE matching, which is suboptimal for abstractive summarization. In this paper, we extend the concept of domain to encompass more refined human preferences, incorporating both ROUGE-based extractive methods and LLM-based abstractive methods to construct supervised training data, which introduces greater diversity and consistently enhances summarization performance.

6 Conclusion

We propose a novel human summarization preference alignment framework ALIGNSUM including Data Pymarid, Gaussian Resampling, and Two-stage Hierarchical Fine-Tuning to align PLMs with human preference. Experiments demonstrate the effectiveness of our framework and narrow the gaps between automatic and human evaluation of PLMs.

Limitations

Dataset Diversity. High-quality preference data acquisition is challenging due to the need for specialized and uniform annotation protocols, along with significant labor and time costs. These preferences are typically implicit and often reflect differences in writing styles, which complicates the annotation process.

Due to the scarcity of preference data, our experiments are limited to *CNN/DailyMail* and *BBC XSum* datasets, as they are the only twos with rewritten versions that reflect human preferences. However, this does not imply that our method is restricted to these datasets. If more high-quality preference datasets become available in the future, we are eager to extend our method to a broader range of datasets.

Language Model Usage. The zero-shot summarization capabilities of LLMs have shown impressive results, making them a seemingly ideal choice for generating summaries that align with human preferences. However, human summarization preferences are inherently implicit, requiring the design of extremely sophisticated prompts to elicit the desired responses from LLMs. This process is challenging and often uncontrollable in real-world scenarios. At the same time, in-context learning is also unrealistic because the text length of the summarization task is much longer than other natu-

ral language tasks, and the upper limit of length is unpredictable.

In contrast, directly fitting implicit preferences using PLMs is a more efficient approach. This method offers irreplaceable advantages in terms of cost and resource consumption, making it a more practical solution for the summarization preference alignment.

Ethics Statement

We utilize publicly available datasets and weight parameters for model training and data generation, all of which are accompanied by bibliographic citations, ensuring no ethical issues are involved.

Acknowledgements

I would like to express my gratitude to the anonymous reviewers for their meticulous and diligent review efforts. This work is supported by the National Science and Technology Major Project 2023ZD0120703 and the China NSFC Projects (U23B2057, 62106142, 62176153 and 62120106006) and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2021. Aspectnews: Aspect-oriented summarization of news documents. *arXiv preprint arXiv:2110.08296*.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2024. Medvoc: Vocabulary adaptation for fine-tuning pre-trained language models on medical text summarization. *arXiv preprint arXiv:2405.04163*.

Guangsheng Bao, Zebin Ou, and Yue Zhang. 2023. Gemini: Controlling the sentence-level summary style in abstractive text summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 831–842.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marc Brysbaert. 2019. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of memory and language*, 109:104047.

Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2023. Unisumm and summzoo: Unified model and diverse benchmark for few-shot summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12833–12855.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Cheng Deng, Tianhang Zhang, Zhongmou He, Yi Xu, Qiyuan Chen, Yuanyuan Shi, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, Zhouhan Lin, and Junxian He. 2023. K2: A foundation language model for geoscience knowledge understanding and utilization.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, SHUM KaShun, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*.

Alexander R Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2020. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. *arXiv preprint arXiv:2010.12836*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Angeliki Gleni, Emmanouil Ktistakis, Miltiadis K Tsilimbaris, Panagiotis Simos, Susanne Trauzettel-Klosinski, and Sotiris Plainis. 2019. Assessing variability in reading performance with the new greek standardized reading speed texts (irest). *Optometry and Vision Science*, 96(10):761–767.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Harold D Lasswell. 1948. The structure and function of communication in society. *The communication of ideas*, 37(1):136–139.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804*.

Yixin Liu, Kejian Shi, Katherine S He, Longtian Ye, Alexander R Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023b. On learning to summarize with large language models as references. *arXiv preprint arXiv:2305.14239*.

Inderjeet Mani and Mark T Maybury. 1999. *Advances in automatic text summarization*. MIT press.

MetaAI. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *Mining text data*, pages 43–76.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Lars St, Svante Wold, et al. 1989. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Yiming Wang, Qianren Mao, Junnan Liu, Weifeng Jiang, Hongdong Zhu, and Jianxin Li. 2022. Noise-injected consistency training and entropy-constrained pseudo labeling for semi-supervised extractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6447–6456.

Yiming Wang, Yuxuan Song, Minkai Xu, Rui Wang, Hao Zhou, and Weiying Ma. 2023b. Retrodiff: Retrosynthesis as multi-stage distribution interpolation. *arXiv preprint arXiv:2311.14077*.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023c. 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.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. Adaptsum: Towards low-resource domain adaptation for abstractive summarization. *arXiv preprint arXiv:2103.11332*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Diffusum: Generation enhanced extractive summarization with diffusion. *arXiv preprint arXiv:2305.01735*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023b. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023c. Summit: Iterative text summarization via chatgpt. *arXiv preprint arXiv:2305.14835*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023d. MACSum: Controllable Summarization with Mixed Attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Xin Chen, and Kai Yu. 2024. Chemdfm: Dialogue foundation model for chemistry.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2020. A joint sentence scoring and selection framework for neural extractive document summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:671–681.

A Detailed Theoretical Derivation: Why Hierarchical Fine-tuning?

According to the Assumption 2.1 in main text, we have

$$
\begin{aligned}
& G_{hy} - G_{hi} = \\
& |H(Y,Z|J(p_{x,y,z;\theta})) - H(Y,Z|J(p_{x;\theta}))| \\
& - |H(Y|J(p_{x,y;\theta})) - H(Y|J(p_{x;\theta}))| \\
& - |H(Z|J(p_{x,y,z;\theta})) - H(Z|J(p_{x,y;\theta}))| \\
= & H(Y,Z|J(p_{x;\theta})) - H(Y,Z|J(p_{x,y,z;\theta})) \\
& + H(Y|J(p_{x,y;\theta})) - H(Y|J(p_{x;\theta})) \\
& + H(Z|J(p_{x,y,z;\theta})) - H(Z|J(p_{x,y;\theta})) \\
= & [H(Y,Z|J(p_{x;\theta})) - H(Y|J(p_{x;\theta}))] \\
& + [H(Y|J(p_{x;\theta})) - H(Z|J(p_{x,y;\theta}))] \\
& - [H(Y,Z|J(p_{x,y,z;\theta})) - H(Z|J(p_{x,y,z;\theta}))] \\
= & [H(Z|Y,J(p_{x;\theta})) - H(Z|J(p_{x,y;\theta}))] \\
& + [H(Y|J(p_{x,y;\theta})) - H(Y|Z,J(p_{x,y,z;\theta}))] \\
= & H(Y|J(p_{x,y;\theta})) - H(Y|Z,J(p_{x,y,z;\theta})) \\
< & 0
\end{aligned}
\tag{8}
$$

B Detailed Experimental Setup

B.1 Human Preference Features of Element-Aware Dataset

The annotators are required to adhere to two types of preferences (Wang et al., 2023c) when writing.

Macro Preference. All news summaries must focus on the four dimensions: **Fluency, Coherence, Consistency, and Relevance**.

Micro Preference. All news summaries should have four essential core elements — **Entity, Date, Event, and Result** — following the "Lasswell Communication Model" (Lasswell, 1948). These elements must be faithful to the source document.

These preferences reflect professional implicit writing styles that are embedded within the text and are difficult to capture explicitly.

B.2 LLM Inference Setting

The GPT-3 results are adapted from SumCoT (Wang et al., 2023c) and reevaluated using the `evaluate` package[5]. GPT-3.5 results are obtained via the OpenAI API with a temperature of 1. The LLaMA series results are inferred from pre-trained weights with a temperature of 0.6.

[5]https://github.com/huggingface/evaluate

B.3 Dataset Examples

We present examples of ED, AD, and HD in *CNN/DailyMail* (Table 18) and *BBC XSum* (Table 17) datasets.

B.4 Main Experiment Packages

Table 9 shows the links of pre-trained model weights and evaluation metrics used in this paper.

Model	URL
BART-Large	https://huggingface.co/facebook/bart-large-cnn https://huggingface.co/facebook/bart-large-xsum
BART-base	https://huggingface.co/ainize/bart-base-cnn https://huggingface.co/Vexemous/bart-base-finetuned-xsum
T5-Large	https://huggingface.co/kssteven/T5-large-cnndm https://huggingface.co/kssteven/T5-large-xsum
PEGASUS	https://huggingface.co/google/pegasus-cnn_dailymail https://huggingface.co/google/pegasus-xsum
LLaMA2	https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
LLaMA3	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Rouge-1/2/L	https://huggingface.co/docs/evaluate/index
BERTScore	https://github.com/Tiiiger/bert_score

Table 9: Links of pre-trained model weights and evaluation metrics used in the paper.

B.5 Selection of LLMs for AD Generation

In this paper, we use LLaMA-2-7B for generating AD, as Table 10 shows minimal improvement when using different LLMs, making LLaMA-2-7B a more efficient choice.

Model	ROUGE-1	ROUGE-2	ROUGE-L
LLaMA2-7B	48.39	23.33	34.47
LLaMA3-8B	48.47	22.83	33.15
LLaMA3-70B	48.92	23.58	34.32

Table 10: Ablation study of using different LLMs to generate AD in ALIGNSUM.

C Reference-free Human Evaluation

We recruit the same annotators (one Ph.D. student and two Master's students) to evaluate 25 randomly selected *BBC XSum* samples based on four criteria: coherence, consistency, fluency, and relevance (rated 1-5, Fabbri et al. (2021)). We select GPT-3, GPT-3 (*w/SumCoT*) and the original summary as baselines, for each sample, we ask the annotators to score from four aspects, and average their rating as the final score. Table 11 presents the average score of the 25 randomly selected samples.

GPT-3 (*w/SumCoT*) outperforms ALIGNSUM in a reference-free setting, but this does not contradict

Method	Coherence	Consistency	Fluency	Relevance
Original Summary	3.4	3.2	4.2	3.16
ALIGNSUM	4.0	3.72	4.48	3.88
GPT-3	4.06	3.6	4.44	3.8
GPT-3 (*w/SumCoT*)	4.32	4.4	4.52	3.96

Table 11: Reference-free human evaluation results across different methods.

our conclusions, as our goal is to align with the specific human preferences reflected in the human-annotated data. On the other hand, ALIGNSUM outperforms the original summary and performs comparably to GPT-3, demonstrating its ability to generate high-quality summaries while aligning with human preferences.

D Supplementary Experimental Results

D.1 BART(w/ full DP) Results on Random Samples

We randomly split the Element-Aware Dataset five times, Table 12 and Table 13 show the automatic evaluation on *CNN/DailyMail* and *BBC XSum*, respectively.

Exp	R1	R2	RL	BERTScore
random1	47.82	22.34	33.02	0.9050
random2	51.82	29.44	38.04	0.9124
random3	48.08	23.24	33.81	0.9045
random4	48.28	22.98	32.88	0.9036
random5	48.17	22.58	33.06	0.9033
Mean	48.83	24.12	34.16	0.9058

Table 12: Experiments on *CNN/DailyMail*

Exp	R1	R2	RL	BERTScore
random1	45.69	20.71	34.42	0.8992
random2	43.25	18.64	32.22	0.8972
random3	41.34	16.88	32.12	0.8959
random4	41.25	16.26	29.58	0.8943
random5	40.35	16.26	29.84	0.8944
Mean	42.38	17.75	31.64	0.8962

Table 13: Experiments on *BBC XSum*

D.2 Case Study

We show the "Win", "Equal" and "Fail" case on *CNN/DailyMail* in Table 14, Table 15 and Table 16 respectively.

BART(w/ full DP)	An elderly woman's body was found mummified in a house in San Francisco's Richmond area, believed to be owned by a hoarder. The body was discovered on Saturday, and police are investigating how it got to this point.
Golden Reference	In one of the worst hoarding cases in recent years, San Francisco police found the body of an elderly woman in a hoarder's home on Saturday, with trash debris piled up in the room. Police are working to investigate the cause of the case.
GPT-3 w/CoT	Police in San Francisco have discovered the mummified body of an elderly woman in a house said to be owned by a hoarder. The property in the Richmond area is said to have been crawling with rats, black widow spiders, dog feces and 300 bottles of urine. The remains, which were removed from the house on Saturday, are believed to be that of Anna Ragin who lived with her 65-year-old daughter Carolyn. Neighbors told the San Francisco Chronicle the mother had not been seen for years. According to records another man, Archibald Ragin, lived at the property but died 15 years ago at the age of 101. Carolyn, said to be a retired Pacific Bell worker, was taken to hospital after her mother's body was removed from the house. Supervisor Eric Mar, a representative in charge of the area, told the Chronicle: T̈he police captains I've spoken with tell me this is the worst case of hoarding they have ever seen. Ḧe was concerned for the health and safety of the residents but questioned how the situation in the property could have gotten to such a point. When city workers arrived to clean up the body on Wednesday, they reportedly struggled to get in the door because the house was filled with so much debris. The mess was said to have included mold and piles of trash that reached the ceiling. They were forced to call in professional workers. The Chronicle quotes a firefighter saying the body was lightweight and wrapped in a blanket. It is not clear where it was discovered.

Table 14: The "Win" Case. BART(w/ full DP) generates a summary of similar length compared to the golden reference, whereas GPT-3 w/CoT produces an excessively long summary.

BART(w/ full DP)	Jason Rezaian, a Washington Post bureau chief in Tehran, has been held in jail in Iran for nearly nine months on unspecified charges of economic spying. His wife, Yeganeh Salehi, was released on bail in October, but he has been denied access to proper legal representation and has been subjected to long interrogations.
Golden Reference	Police arrested Jason Rezaian, the Washington Post's Tehran bureau chief, on July 22 on unknown charges, and did not make the charges public until last week, more than four months after the arrest.
GPT-3 w/CoT	Jason Rezaian, the Washington Post's bureau chief in Tehran, has been in jail in Iran for nearly nine months on unspecified charges. Last week, it was revealed that he will be tried soon on espionage charges. The Washington Post has condemned the charges as äbsurdänd called for Rezaian's release. Rezaian has been denied bail and has been subjected to long interrogations. His wife was released on bail in October.

Table 15: The "Equal" case. We represent the entities, dates, events, and results in red, blue, green and yellow. It is clear that $\text{Info}_1 = \text{Info}_2 = 3$.

BART(w/ full DP)	On Wednesday, Andrea Atzeni will ride terror in the 1,000 Guineas at Newmarket, who finished fourth in the group one Cheveley park stakes over six furlongs at the Rowley mile last September. David Simcock, trainer of terror, believes she has been solid in training and is not a bad filly, but it's difficult to know where she stands fitness-wise until she gets into a race scenario.
Golden Reference	Jockey Andrea Atzeni will ride Terror in Wednesday's Landwades Nell Gwyn Stakes at Newmarket. Terror, trained by David Simcock, is one of the contenders for the 1,000 Guineas. This race will be a great opportunity for Atzeni to publicize the horse's chances in the Classic.
GPT-3 w/CoT	Jockey Andrea Atzeni will ride the horse Terror in the Landwades Nell Gwyn Stakes at Newmarket on Wednesday. Terror is trained by David Simcock and is one of the contenders for the 1,000 Guineas. This race will be a good opportunity for Atzeni to promote the horse's chances in the classic.

Table 16: The "Fail" case. We represent the entities, dates, events, and results in red, blue, green and yellow. It is clear that $\text{Info}_1 = 4 < \text{Info}_2 = 7$.

Document	conrad clitheroe and gary cooper , both from stockport , and expat neil munro were reportedly taking notes near fujairah airport , 80 miles from dubai , when they were arrested in february . relatives were told they were held for " national security " reasons . the men insisted they did not take photographs . the abu dhabi hearing is due on monday . mr clitheroe , 54 , and mr cooper 45 , were visiting their friend mr munro , who was born in manchester , when they were arrested on 22 february by an off-duty police officer who had seen them monitoring planes from a car . they were near fujairah airport , where older and rarer aircraft can be seen . a local police official said the men had been taking photographs near an airport and were using a telescope . the men are expected to argue their actions were misinterpreted and are understood to be hoping to be granted bail .
ED	mr clitheroe , 54 , and mr cooper 45 , were visiting their friend mr munro , who was born in manchester , when they were arrested on 22 february by an off-duty police officer who had seen them monitoring planes from a car .
AD	Three British men, Conrad Clitheroe and Gary Cooper from Stockport and Neil Munro from Manchester, were arrested near Fujairah Airport in February for taking notes and using a telescope, with their lawyer expected to argue that their actions were misinterpreted and they are hoping to be granted bail.
HD	Three men were arrested for taking notes and taking photographs near fujairah airport in February, they hope to be granted bail for being misinterpreted.

Table 17: Case of ED, AD and HD in *BBC XSum*

Article	A married software executive who drugged a female employee in order to take naked pictures of her on a business trip has been jailed. Sexual predator Henri Morris was told he would serve 10 years behind bars for his 'calculated and choreographed' crime. The 67-year-old was caught in an FBI sting after investigators were approached by one of his victims in 2012. Henri Morris, 67, was jailed for 10 years after admitting drugging a female employee during a business trip in order to take naked photos of her She told them that her drink was spiked by the married businessman after they traveled together from Houston, Texas, to New Jersey for work. The woman said when she woke up she was naked and her boss was standing over her and taking pictures on his mobile phone. The FBI arrested Morris at Bush County Airport after the woman, who has not been named, covertly worked with them. When his bags were search they found his 'kit,' which included strong sedatives and Viagra. A task force was set up to probe Morris and in total they found eight female employees who claimed to have been abused by him. His defense team initially tried to argue that the women all willingly drank to excess with clients and that no abuse had ever occurred. Ahead of his trial Morris pleaded guilty to a single charge of drugging and abusing a female employee, in exchange for prosecutors dropping other abuse charges. MarriedMorris of Houston, Texas, was the president of Edible Software Solutions 'There is no way to adequately express my remorse and my abject humiliation,' Morris said during a 10 minute address at his sentencing hearing, according to the Houston Chronicle. 'I apologize to anyone who I have hurt and beg for their forgiveness.' US District Judge Melinda Harmon sentenced him to 10 years in federal prison without the possibility of parole. She also reportedly ordered that he spend the rest of his life under the supervision of the US Bureau of Prisons, according to the newspaper.
ED	Henri Morris, 67, was jailed for 10 years after admitting drugging a female employee during a business trip in order to take naked photos of her She told them that her drink was spiked by the married businessman after they traveled together from Houston, Texas, to New Jersey for work.
AD	A married software executive, Henri Morris, has been sentenced to 10 years in federal prison for drugging and sexually assaulting a female employee during a business trip. Morris spiked the woman's drink with sedatives and took naked pictures of her while she was unconscious. The FBI arrested Morris after one of his victims came forward and the task force found eight other female employees who had been abused by him.
HD	Henri Morris, 67, was a married software executive, he was jailed for 10 years for drugging a female employee during a business trip in order to take naked photos of her, without the possibility of parole.

Table 18: Case of ED, AD and HD in *CNN/DailyMail*