Learning to Summarize from LLM-generated Feedback

Anonymous ACL submission

Abstract

Developing effective text summarizers remains a challenge due to issues like hallucinations, key information omissions, and verbosity in LLM-generated summaries. This work explores using LLM-generated feedback to improve summary quality by aligning the summaries with human preferences for faithfulness, completeness, and conciseness. We introduce FeedSum, a large-scale dataset containing multi-dimensional LLM feedback on summaries of varying quality across diverse domains. Our experiments show how feedback quality, dimensionality, and granularity influence preference learning, revealing that high-quality, multi-dimensional, fine-grained feedback significantly improves summary generation. We also compare two methods for using this feedback: supervised fine-tuning and direct preference optimization. Finally, we introduce SummLlama3-8b, a model that outperforms the nearly 10x larger Llama3-70b-instruct in generating human-preferred summaries, demonstrating that smaller models can achieve superior performance with appropriate training. The full dataset and model will be released upon acceptance. A portion of the dataset is available beforehand at `https://bit.ly/4eGdVFw`.

1 Introduction

Developing an effective text summarizer has long been a challenge, as summaries generated by language models often fall short of human standards (Maynez et al., 2020; El-Kassas et al., 2021; Roit et al., 2023). While large language models (LLMs) have greatly improved the coherence and fluency of summaries (Liu et al., 2023a), persistent issues remain, such as unfaithful statements (hallucinations), omission of key information (low completeness), and verbosity (low conciseness) in the summaries (Lee et al., 2024; Song et al., 2024).

An ideal approach would involve providing expert-level summary examples for language models to imitate. However, creating such high-quality summaries is both labor-intensive and difficult to scale effectively. A better alternative is to leverage feedback on the summaries through reinforcement learning from human feedback (RLHF) (Stiennon et al., 2020; Rafailov et al., 2023). RLHF avoids the need to write an ideal summary by having users select their preferred response from candidate summaries of the same document. Yet, human involvement still poses scalability challenges, particularly when acquiring high-quality, fine-grained human feedback across multiple dimensions, such as faithfulness and completeness. For example, Lee et al. (2024) reports that the cost of obtaining fine-grained human feedback for these three dimensions exceeded $30K for just 2,025 summaries.

In this work, we address this challenge by utilizing *LLM-generated* feedback, known as RL from AI feedback (RLAIF) (Lee et al., 2023; Dutta et al., 2024), with a special focus on text summarization. Our approach shifts focus to the relatively unexplored area of leveraging LLM-generated feedback to enhance summary quality, whereas most existing research in summarization has primarily concentrated on using LLMs to evaluate summaries. (Wan et al., 2024; Tang et al., 2024a; Song et al., 2024). Specifically, our goal is to produce human-preferred summaries by exploiting LLM feedback with respect to the three core dimensions, namely *faithfulness*, ensuring summaries are consistent with original documents; *completeness*, encompassing all key-facts[1]; and *conciseness*, maintaining a succinct and focused summary. We focus on these three dimensions, as LLMs already excel in other axes like fluency and coherence (Liu et al., 2023a; Song et al., 2024). Figure 1 illustrates our complete pipeline for learning from LLM-generated feedback, consisting of data sourcing, feedback generation, and preference learning.

[1]A key-fact is a concise sentence that conveys a single key piece of information (Bhandari et al., 2020; Song et al., 2024).

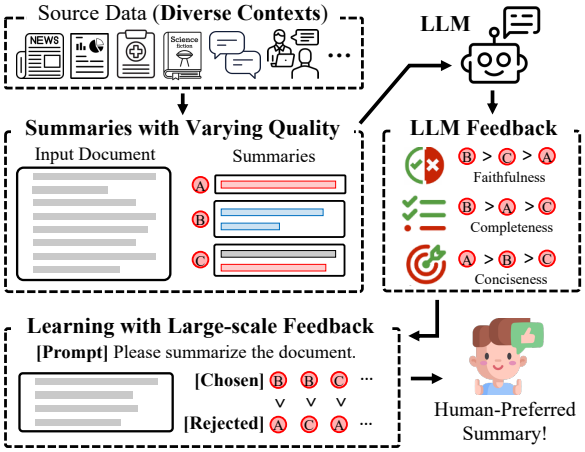


Figure 1: FeedSum consists of summaries of varying quality, generated by 13 different summarizers across input documents from 7 distinct domains. Through automated evaluation using LLMs, 125K document-summary pairs have been produced, each accompanied by LLM-generated multi-dimensional feedback, providing valuable data for preference learning.

We begin by creating a large-scale dataset, **FeedSum**, with multi-dimensional LLM feedback on text summaries. To obtain summaries with varying levels of quality, we generate them using 13 different language models, including 3 non-LLMs (e.g., Bart), 7 open-source LLMs (e.g., Llama3), and 3 proprietary LLMs (e.g., GPT-4-turbo). For summary generation, these models are applied to a diverse range of input documents, spanning both short and lengthy texts, including dialogue and non-dialogue formats, and across 7 distinct domains.

The effectiveness of LLM feedback on preference learning can vary based on three factors, including the *quality* of the feedback (e.g., low vs. high), its *dimensionality* (e.g., single vs. multiple), and its level of *granularity* in scoring (e.g., coarse vs. fine). To explore the impact of these factors, we configure four different setups (see Table 2 for details). Each setup generates 125K pairs of input documents and summaries, accompanied by distinct LLM-generated feedback responses. In particular, we find that these factors significantly affect the distribution of feedback scores, which is crucial for determining "chosen" versus "rejected" summaries in preference learning.

Through extensive experiments with FeedSum, we provide insights into learning to summarize from LLM feedback, focusing on three key open questions: *(Q1)* The impact of these three factors on the effectiveness of preference learning; *(Q2)* An analysis on the effectiveness of each dimension in multi-dimensional feedback; and *(Q3)* A comparison between two approaches for utilizing LLM-generated feedback: supervised fine-tuning (SFT) and direct preference optimization (DPO).

Our main contributions are: (1) We create and release FeedSum, the first large-scale summarization dataset for preference learning, featuring high diversity in inputs, summaries, and feedback; (2) We examine how different configurations of LLM-generated feedback impact preference learning, and the importance of feedback quality, dimensionality, and granularity. (3) We examine the alignment trade-off associated with relying on a single dimension for preference learning. (4) We compare the effectiveness of DPO and SFT variants. (5) We plan to release SummLlama3-8b, which outperforms the nearly 10x larger Llama3-70b-instruct in the three human-preferred dimensions.

2 Related Work

Preference Optimization. Preference optimization plays a crucial role in bridging the gap between human intent and the outputs generated by LLMs (Yang et al., 2023; Jiang et al., 2024b; Rafailov et al., 2023). The predominant methods are PPO (Schulman et al., 2017)[2], which fits a reward model to optimize LLMs to generate responses that receive high reward, and DPO (Rafailov et al., 2023), which directly optimizes the LLMs' outputs based on preference data without relying on an explicit reward model. These methods have demonstrated effectiveness in aligning LLMs with human preferences, particularly in mitigating hallucination, harmful outputs, and biased contents (Tonmoy et al., 2024; Bai et al., 2024; Allam, 2024; Li et al., 2024). Despite the success of preference optimization in other domains, in text summarization, limited work has focused on aligning outputs with human preferences. Stiennon et al. (2020) collected a comparison dataset to train a PPO reward model, using Reddit posts and coarse evaluations of two summaries, without accounting for multi-dimensional aspects of summarization. Recently, Mishra et al. (2024) applied DPO for summarization alignment, but focusing solely on faithfulness using low-quality summaries by synthetically altering high-quality ones with LLMs.

Automated Evaluation. Conventional metrics, such as ROUGE and BERTScore, have shown

[2]PPO is not well-suited for tasks with lengthy input and output, such as text summarization, due to its iterative auto-regressive generation of summaries at each training iteration. Therefore, we primarily focus on using DPO.

Dataset	# of Domain	Input Type	Input Length	Summary Gen.	# of Feedback Dim.	Feedback Size	Feedback Type
UniSumEval	Multiple (9)	Dial. & Non-dial.	Short–Long	Summarizers (9)	Faith, Comp, Cons	1K	Human-annotated
SynFacEdit	Single (1)	Non-dial.	Short	Synthetic Edit (1)	Faith	5K	LLM-generated
FeedSum (Ours)	Multiple (7)	Dial. & Non-dial.	Short–Long	Summarizers (13)	Faith, Comp, Cons	125K	LLM-generated

Table 1: Comparison of FeedSum with existing summarization datasets with human feedback, UniSumEval (Lee et al., 2024), and LLM feedback, SynFacEdit (Mishra et al., 2024). The numbers in the parenthesis are the number of input domains (in "# of Domain") and summarizers used for summary generation (in "Summary Gen").

a weak correlation with human judgment in the evaluation of text summaries (Pagnoni et al., 2021; Tang et al., 2024b; Song et al., 2024). In response, several methods have emerged, focusing on fine-tuning models using well-curated datasets, including natural language inference (NLI)-based and question answering (QA)-based approaches (Fabbri et al., 2022; Laban et al., 2022; Zhong et al., 2022). However, they are typically limited to evaluating only the faithfulness of summaries and require training specialized models. The emergence of large language models (LLMs) facilitates multi-dimensional evaluation by incorporating them into human-like evaluation pipelines (Wang et al., 2023; Tang et al., 2024a; van Schaik and Pugh, 2024; Fu et al., 2024). Particularly, G-Eval (Liu et al., 2023a) uses GPT-4 for automated evaluation, providing 1-5 Likert-scale scores based on the prompt. FineSurE (Song et al., 2024) adopts fine-grained protocols for sentence-level faithfulness and key-fact-level completeness and conciseness, producing objective percentage scores, such as the proportion of faithful sentences and key-facts included.

In this work, we integrate automated evaluation into preference learning, enabling large-scale, fine-grained feedback that addresses three key alignment dimensions of text summarization: faithfulness, completeness, and conciseness.

3 Data and Experiment Details

High-level Overview. We overview the overall pipeline from data creation to preference learning, as in Figure 1, following three distinct steps:

Step 1. Input sourcing and summary generation: Input documents are extracted from 7 diverse sources, varying in domain, length, and type. Summaries are then generated using 13 non-LLMs, open-source LLMs, and proprietary LLMs, producing a wide range of summary quality.

Step 2. Feedback generation using LLMs: Feedback is generated through LLM-based summary evaluations using four configurations, adjusting the quality (low vs. high), dimensionality (single vs. multi-dimensional), and granularity (coarse- vs. fine-grained) of the feedback.

Step 3. Learning with large-scale feedback: We examine the potential of machine feedback from LLMs through the lens of preference learning.

3.1 FeedSum: Data with LLM Feedback

Table 1 contrasts our FeedSum dataset with two existing datasets with either human-annotated or LLM-generated feedback. FeedSum features input text that is as diverse as UniSumEval (Lee et al., 2024), while simultaneously providing a significantly larger amount of LLM-generated feedback compared to SynFacEdit (Mishra et al., 2024).

Input Text Sourcing The diversity of source documents is crucial for acquiring comprehensive feedback, as it helps identify weaknesses in modern text summarizers across various aspects such as input domain, length, and type (dialogue vs. non-dialogue) (Lee et al., 2024). Hence, we extract input documents from multiple source datasets encompassing 7 different domains, ranging from short to lengthy input texts, and covering both non-dialogue and dialogue formats. We sample 2,000 input texts from the training set of each source dataset, including four non-dialogue datasets – CNN/DM (news) (Nallapati et al., 2016), Wikihow (lifestyle) (Koupaee and Wang, 2018), GovReport (report) (Huang et al., 2021), PubMed (medical literature) (Cohan et al., 2018) – and three dialogue datasets – DialogSum (dailylife) (Chen et al., 2021), MediaSum (interview) (Zhu et al., 2021), MeetingBank (meeting) (Hu et al., 2023). As a result, a total of 14K input documents are sampled.

Summaries with Varying Quality. The performance of summarization can vary depending on the summarizer chosen, as there is no single model that consistently produces the best quality summary (Song et al., 2024). From the perspective of preference learning, it is crucial to gather feedback from a wide range of summaries with varying quality. This allows for pairwise comparisons of different quality summaries even for the same document. Therefore, we choose 13 language models grouped into

ID	Quality	Dimension	Granularity	Feedback Example
C1	Low (Llama3-8b)	Single	Summary-level	{Overall Score: 4}
C2	High (Llama3-70b)	Single	Summary-level	{Overall Score: 5}
C3	High (Llama3-70b)	Multiple	Summary-level	{Faithfulness: 5, Completeness: 3, Conciseness: 3}
C4	High (Llama3-70b)	Multiple	Sentence & key-fact	{Faithfulness: 75%, Completeness: 66%, Conciseness: 25%}

Table 2: Four different configurations to acquire LLM-generated feedback on the quality of text summaries.

three categories, namely *non-LLMs*, including Bart-large (Lewis et al., 2020), Pegasus-large (Zhang et al., 2020), and Flan-t5-large (Chung et al., 2024), *open-source LLMs*, including Phi-2 (Javaheripi et al., 2023), Mistral-7b-instruct (Jiang et al., 2023), Mixtral-8x7b-instruct (Jiang et al., 2024a), Llama2-13b-chat (Touvron et al., 2023), Mistral-nemo-instruct, Llama3-70b-instruct, and Gemma2-27b-instruct (Team et al., 2024), and *proprietary LLMs*, including Claude-instant, GPT-3.5-turbo, and GPT-4-turbo (Achiam et al., 2023). These summarizers generate rich, diverse summaries, producing 182K document–summary pairs.

3.2 LLM-generated Feedback

We generate feedback by conducting automated evaluations using LLMs. Yet, the effectiveness of LLM feedback varies the evaluation configuration employed. This study addresses open questions about leveraging LLM-based feedback for preference learning, with a focus on its "quality," "dimensionality," and "granularity." These aspects are investigated by contrasting the effectiveness of feedback generated from four distinct configurations (C1–C4) in Table 2, as summarized below:

• **Feedback Quality (C1 vs. C2)**: The quality of generated feedback plays a pivotal role in preference learning. To assess the importance of feedback quality, we adjust the capacity of the selected LLMs for feedback generation. We use two open-source LLMs of different sizes: Llama3-8b-instruct for low-quality feedback and Llama3-70b-instruct for high-quality feedback, respectively.

• **Feedback Dimensionality (C2 vs. C3)**: The simplest way to gather feedback is to assess the quality of the summary with a single score on a 1–5 Likert scale (Wang et al., 2023). However, it overlooks key multi-dimensional aspects of summary quality, such as faithfulness, completeness, and conciseness (Lee et al., 2024). Therefore, a more advanced approach involves conducting a multi-dimensional evaluation using LLMs across these three dimensions, yielding a separate score for each (Zhong et al., 2022; Liu et al., 2023a).

• **Feedback Granularity (C3 vs. C4)**: Coarse-grained evaluation uses a Likert scale (1–5), but these subjective scores often skew toward higher ratings due to a lack of grounding (Wang et al., 2023; Liu et al., 2023a). In contrast, fine-grained evaluation assesses at the sentence or key-fact level, measuring faithfulness, completeness, and conciseness through factual sentence proportions and key-fact coverage, yielding percentage scores better correlate with human feedback (Song et al., 2024).

Thus, all summary–document pairs are subjected to LLM-based summarization evaluation for each configuration. 125K summary-document pairs remain after excluding failed and special cases.[2] More details including evaluation prompts, methodologies, and data statistics are in Appendix A.

3.3 Learning with Large-scale Feedback

We primarily focus on three key questions:

Q1: How do the quality, dimensionality, and granularity of LLM-generated feedback (C1–C4 in Table 2) influence preference learning?

Q2: What impact does each dimension have in the case of multi-dimensional feedback?

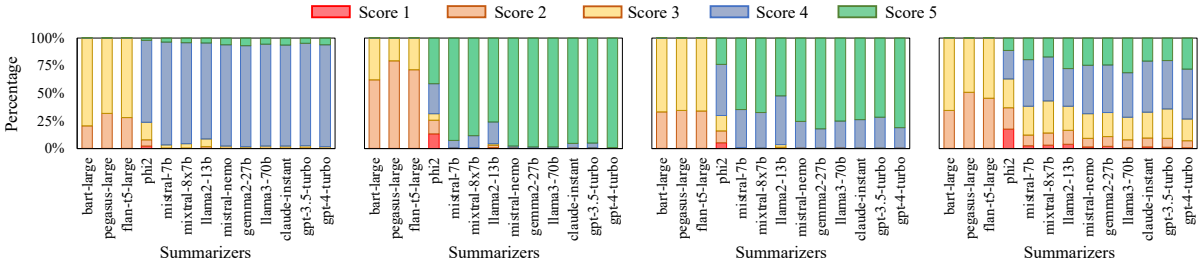
Q3: How much can DPO enhance the quality of summaries compared to SFT variants?

The experimental setups are detailed below and see Appendix B for the detailed training configuration and input–output format for SFT and DPO.

Q1: Impact of Feedback Configuration. In this experiment, we consistently train Llama3-8b-instruct using DPO (Rafailov et al., 2023), but with feedback generated by differently configured LLMs (C1–C4 in Table 2). Specifically, for each configuration, we create a set of paired summaries – one chosen and one rejected – using the same criteria to determine their selection. A summary is selected as the "chosen" one if it scores ≥ 4 on the Likert scale or $\geq 80\%$ in percentage scores[3].

[2] We exclude cases where the document length exceeds the model's capacity (8K tokens by Llama3 on a single NVIDIA H100 GPU), as well as instances of erroneous feedback from the LLMs, such as incomplete or nonsensical responses.

[3] For multi-dimension cases (C3 and C4), we use a composite score, the average across all dimensions by default.



(a) Configuration 1 (C1). (b) Configuration 2 (C2). (c) Configuration 3 (C3). (d) Configuration 4 (C4).

Figure 2: Distribution of summary scores on a 1–5 Likert scale across the four different configurations. Percentage scores in C4 are converted into Likert-scale ones through uniform quantization for ease of interpretation.

Level	C1	C2	C3	C4
Summary	0.358	0.470	0.589	0.649
System	0.800	0.833	0.783	0.983

Table 3: Quality of feedback using C1–C4 assessed based on their Spearman correlation with human composite scores in UniSumEval (Lee et al., 2024), evaluated at both the summary and system levels.

Type	Non-LLMs	Open. LLMs	Propr. LLMs
C1	10.7 (60.7)	56.3 (29.6)	33.0 (9.7)
C2	6.2 (74.4)	59.4 (23.3)	34.4 (2.3)
C3	5.1 (81.2)	60.9 (18.5)	34.0 (0.3)
C4	19.3 (30.9)	55.1 (50.2)	25.6 (18.9)

Table 4: Proportion (%) of chosen (and rejected) summaries across three summarizer categories.

Conversely, a summary is considered "rejected" if its score is at least 1 point lower on the Likert scale or 20 percentage points lower than the chosen one. To ensure a fair comparison, we standardize the number of pairs for each configuration to 92K, matching the number provided by C4. For clarity, we refer to the four summarizers as *DPO-C1*, *DPO-C2*, *DPO-C3*, and *DPO-C4*.

Q2: Impact of Feedback Dimension. We examine how focusing on a specific dimension affects summary quality compared to considering all dimensions equally. This experiment provides insights into the concept of alignment tax (Noukhovitch et al., 2023; Guo et al., 2024), which refers to the trade-off where enhancing alignment with one objective (e.g., conciseness) may reduce performance in another (e.g., completeness). For each objective, we generate pairs of chosen and rejected summaries using the same criteria as in Q1. We then fine-tune Llama3-8b-instruct using DPO for preference learning, resulting in three compared summarizers: *DPO-faith*, *DPO-comp*, and *DPO-cons*, each focusing on faithfulness, completeness, and conciseness, respectively.

Q3: Comparison of DPO over SFT. We compare the effectiveness of DPO with that of five SFT variants. The performance of SFT heavily depends on which summaries are used as reference for teacher-forcing (Han et al., 2024). We consider four different setups, namely *SFT-human*: We fine-tune Llama3-8b-instruct using human summaries in our source datasets as reference; *SFT-best*: We first identify the best summary among those generated by 13 summary models in FeedSum, selecting the one with the highest composite score obtained from C4. Then, use it as the reference to supervise Llama3-8b-instruct; *SFT-faith*, *SFT-comp*, and *SFT-cons*: Similar to SFT-best, but the best summary is selected based solely on their faithfulness, completeness, and conciseness scores, respectively.

4 Analysis of LLM-generated Feedback

The configurations C1–C4 for LLM-based feedback generation show significant differences in feedback quality and their impact on constructing chosen-rejected summary pairs.

4.1 Quality of LLM Feedback

Table 3 presents the quality of LLM feedback by calculating summary- and system-levels agreement with human scores of UniSumEval (see Appendix C.2 for the detailed metric). The feedback in C1 shows a significantly lower correlation with human scores compared to C2–C4, which is attributed to the use of a smaller LLM. The quality of feedback improves progressively as it becomes more multi-dimensional and fine-grained from C2 to C4.

4.2 Distribution of Summary Score

Figure 2 illustrates the distribution of composite summary scores obtained from the four LLM-based evaluation strategies (C1–C4), as in Table 2.

Firstly, **low-quality feedback from C1 tends to avoid assigning very low (score 1) or high (score 5) scores.** As shown in Figure 2(a), most

Backbone	Config.	Quality	Dimension	Granularity	Faith.	Comp.	Conc.	Avg.
Llama3-8b-inst.	wo. RL				0.864	0.583	0.450	0.632
Llama3-8b-inst.	DPO-C1	Low	Single	Coarse-grained	0.836	0.594	0.453	0.628
	DPO-C2	High	Single	Coarse-grained	0.878	0.595	0.514	0.662
	DPO-C3	High	Multiple	Coarse-grained	0.889	0.581	0.516	0.662
	DPO-C4	High	Multiple	Fine-grained	**0.931**	**0.614**	**0.659**	**0.735**
Llama3-70b-inst.	wo. RL				0.931	0.596	0.487	0.671
Llama3-70b-inst.	DPO-C4	High	Multiple	Fine-grained	**0.950**	**0.632**	**0.754**	**0.779**

Table 5: Automated evaluation results of seven summarizers on faithfulness, completeness, and conciseness. "w.o RL" refers to the models without preference optimization. "DPO-{C1,C2,C3,C4}" denote models incorporating preference optimization, utilizing feedback generated in C1–C4 of Table 2. The best score are marked in bold.

summaries receive scores between 3 and 4, regardless of the summarizer they were generated from. This results in the incorrect selection of low-quality summaries as the chosen ones.

Secondly, **feedback from the coarse-grained evaluation in C2–C3 introduces a significant bias, favoring LLM-generated summaries over non-LLM ones**, regardless of whether single- or multi-dimensional evaluation is used. Only a few LLM-generated summaries have scores below 4. This leads to the issue of indiscriminately selecting LLM summaries as chosen ones while non-LLM summaries as rejected ones.

Lastly, **feedback from the fine-grained evaluation in C4 is robust to the summarizer category, showing a highly diverse score distribution across all summarizers.** This approach not only accurately captures the hierarchy among non-LLMs, open-source LLMs, and proprietary LLMs, but also demonstrates that high-quality summaries can be produced by older models. That is, summaries generated by LLMs can be rejected in favor of higher-quality ones produced by non-LLMs.

4.3 Chosen and Rejected Summary

Table 4 summarizes the proportion of summaries selected as chosen or rejected ones across the three summarizer categories. **A significant difference in proportion is observed depending on the feedback configuration.** Feedback from single-dimensional or coarse-grained automatic evaluations in C1–C3 predominantly classifies LLM-generated summaries as chosen in 89.3%–94.9% of cases, while non-LLM-generated summaries are rejected in 60.7%–81.2% of cases. However, feedback from the multi-dimensional and fine-grained evaluation in C4 reveals a considerably different trend, with 69.1% of LLM-generated ones as rejected and 19.3% of non-LLM-generated summaries being classified as chosen. This significant difference across configurations greatly affects their effectiveness in preference optimization.

5 Main Experiment

Test Set. The test set is constructed by randomly sampling 200 documents from the test split of FeedSum's seven source datasets, totaling 1.4K. Both *automated* and *human* evaluations are conducted to evaluate summarizers on this set. For automated evaluation, we use Llama3-80b-instruct as the backbone of FineSurE (Song et al., 2024). For human evaluation[3], we perform a fact verification task for faithfulness and a key-fact alignment task for completeness and conciseness, following the work (Lee et al., 2024). Three annotators are assigned for each task, recruited through Amazon Mechanical Turk. Details of the automated and human evaluation can be found in Appendix C.3.

Evaluation Metric. We report the quality of text summary from the three key perspectives: *faithfulness*, the proportion of faithful summary sentences; *completeness*, the proportion of covered key-facts; and *conciseness*, the proportion of summary sentences aligning with the key-facts. In automated evaluation, key-facts are automatically extracted from the reference summary of each source dataset, as suggested by Song et al. (2024). In human evaluation, we use human-annotated key-facts from the UniSumEval dataset (Lee et al., 2024) to follow the same manual evaluation pipeline in the paper. See Appendix C.1 for the equation to calculate three percentage scores. Appendix D presents the results using ROUGE and BERTScore for reference.

5.1 Q1: Impact of Feedback Configuration

We evaluate the quality of summaries generated by Llama3, trained with DPO incorporating feedback

[3]Due to the high cost of fine-grained human evaluation, 420 document-summary pairs are randomly sampled, i.e., 10 (documents) $\times$ 7 (source data) $\times$ 6 (summarizers) in Table 5.

Config.	Faith.	Comp.	Conc.	Avg.
Llama3-8b	0.902	0.636	0.784	0.774
Llama3-70b	0.953	0.659	0.792	0.801
DPO-C1	0.868	0.669	0.826	0.787
DPO-C2	0.947	0.675	0.840	0.820
DPO-C3	0.925	0.664	0.869	0.819
DPO-C4	**0.980**	**0.697**	**0.959**	**0.879**

Table 6: Human evaluation of six summarizers on three dimensions. Llama3-8b-instruct and Llama3-70b-instruct (without RL) are baselines, while C1–C4 are Llama3-8b-instruct after DPO with varying feedback.

Strategy	Faith.	Comp.	Conc.
DPO-faith	**0.942 (+0.078)**	0.577 (-0.006)	0.686 (0.236)
DPO-comp	0.846 (-0.018)	**0.640 (+0.057)**	0.438 (-0.012)
DPO-cons	0.877 (+0.013)	0.493 (-0.090)	**0.892 (+0.442)**
DPO-avg	0.902 (+0.038)	0.608 (+0.025)	0.591 (+0.141)

Table 7: Automated evaluation results of summarizers trained with DPO on a specific dimension. DPO-avg is a model, averaging the LoRA weights of others. Cells are color-coded from dark to light based on descending score ranks for each dimension. The value in the parenthesis is gain or decline over the model wo. DPO.

from four different configurations, and compare the results to the corresponding models without DPO. The prompts used to generate summaries are identical across all setups, as in Appendix E. The results for other LLMs are in Appendix F.

5.1.1 Results by Automated Evaluation

Table 2 presents the summary quality of seven summarizers: two models without preference optimization and five models with DPO using LLM feedback generated in different configurations.

The low-quality feedback (C1) generated by Llama3-8b-instruct proves ineffective. Compared to not using DPO, the faithfulness score rather drops by 0.028, resulting in the lowest composite score ("Avg.") across the three dimensions. The quality of feedback is crucial for preference learning using LLM feedback. **High-quality feedback (C2–C3) via coarse-grained evaluation improves the performance of summarizers in most cases.** However, there is no improvement in considering multi-dimensional aspects in automated evaluation as long as the granularity remains coarse. This is likely due to coarse-grained feedback lacking diversity in feedback pairs, indiscriminately selecting LLM-generated summaries as chosen and non-LLM summaries as rejected, as shown in Figures 2(b) and (c). **To fully exploit LLM-generated feedback, fine-grained evaluation (C4) with diverse score distribution is crucial.** The improvement is significant, achieving a composite score

Strategy	Faith.	Comp	Conc	Abs.	Avg.
Llama3-8b	0.864	0.583	0.450	0.696	0.648
DPO-C4	0.931	0.614	0.659	0.691	0.723
SFT-human	0.774	0.496	**0.666**	0.516	0.613
SFT-best	0.894	0.551	0.572	**0.588**	**0.651**
SFT-faith	**0.903**	0.545	0.536	0.559	0.635
SFT-comp	0.871	**0.597**	0.511	0.594	0.643
SFT-conc	0.874	0.483	0.632	0.488	0.620

Table 8: Comparison of DPO with SFT variants using automated evaluation, where they are fine-tuned from Llama3-8b-instruct. For SFT, the best score is marked in bold, while the second-best score is underlined.

of 0.735, which is 0.103 higher than Llama3-8b-instruct (w.o RL) and even surpasses the nearly 10x larger Llama3-70b-instruct (w.o RL).

Additionally, the improvement by fine-grained feedback (C4) even remains with Llama3-70b-instruct. We name our best models SummLlama3-8b/70b and plan to release them on Huggingface. Examples of summaries from different configurations are compared in Appendix G.

5.1.2 Results by Human Evaluation.

Table 6 presents the results of the human evaluation across three dimensions. The overall performance dominance aligns with the automated evaluation results presented in Table 5. The DPO-C4, which is Llama3-8b-instruct fine-tuned with DPO using feedback from C4, significantly outperforms DPO-{C1, C2, C3}, and even surpasses the larger Llama3-70b-instruct. Therefore, the results from both automated and human evaluations confirm that **a smaller model can outperform its larger counterpart with appropriate training.**

5.2 Q2: Impact of Feedback Dimension

Table 7 shows the summary quality of summarizers trained with DPO on a single feedback dimension from C4, along with the model obtained through post-hoc parameter merging (Jang et al., 2023).

Compared to the original Llama3-8b-instruct, DPO-{faith, comp, cons}, which rely on a single dimension, achieve the best scores in their target dimensions, but **they are likely to show limited improvements or even performance declines in other dimensions**. Specifically, focusing on completeness lowers faithfulness and conciseness, while prioritizing conciseness reduces completeness. Additionally, their parameter merging, DPO-avg, results in a balanced score across all dimensions, but training with chosen-rejected pairs based on the composite scores from the sixth row of Table 5 (DPO-C4) achieves better results.

Dim.	0K	13K	23K	46K	92K
Faith.	0.864	0.914	0.937	0.930	0.931
Comp.	0.583	0.613	0.603	0.612	0.614
Conc.	0.450	0.594	0.611	0.648	0.659
Avg.	0.632	0.707	0.717	0.730	0.735

Table 9: Automated evaluation results as the number of selected-rejected summary pairs of LLM-generated feedback for DPO increases from 0 to 92K.

Source	Size	Faith.	Comp.	Conc.	Avg.
UniSumEval	1K	0.874	0.618	0.506	0.666
SynFacEdit	5K	0.789	0.520	0.563	0.624
FeedSum (Ours)	5K	0.913	0.606	0.587	0.702
	92K	0.931	0.914	0.659	0.735

Table 10: Effectiveness of feedback in FeedSum over UniSumEval (human feedback) and SynFacEdit (synthetic feedback). "Size" refers to the available number of pairs consisting of chosen and rejected summaries.

5.3 Q3: Comparison of DPO over SFT

Table 8 compares the summary quality of five SFT variants fine-tuned with reference summaries selected based on different policy, alongside Llama3-8b-instruct before and after DPO. Here, we introduce an additional dimension, "abstractiveness (Abs.)," which refers to the extent to which a summary generates novel sentences or phrases, leading to a more coherent summary (Zhang et al., 2022; Song et al., 2023). The quality of generated summaries vary depending on the SFT policy.

Firstly, **DPO is a much superior approach to the SFT variants.** DPO-C4 significantly improves summary quality across multiple dimensions when compared to the vanilla Llama3-8b-instruct, while SFT-best falls short in delivering similar improvements, despite being fine-tuned with only the best-selected summaries during the training phase. Secondly, **all SFT variants show a notable decline in abstractiveness**, incurring the *copy bias* and leading to less coherent summaries due to sentence copying from the input document (Song et al., 2023). This is because SFT allows only one reference summary per document, whereas DPO is superior by presenting multiple possible summaries through chosen-rejected pairs. Thirdly, **focusing on a single dimension in SFT may improve that aspect but is likely to worsen others.** SFT-faith and SFT-cons improve their target dimensions compared to Llama3-8b-instruct but both sacrifice completeness. Thus, SFT-best, which equally considers all dimensions, achieves the highest average score ("Avg.") among the SFT variants.

5.4 Additional Experiment

Feedback Size. We explore the impact of varying the size of high-quality, multi-dimensional, fine-grained feedback generated by LLMs in C4. The percentage scores, based on the number of selected-rejected pairs used for DPO, are presented in Table 9. Overall, the scores gradually improve as the amount of feedback pairs increases. Notably, there is a significant improvement with as few as 13K summary feedback pairs, but the increase nearly plateaus after 46K feedback pairs. Thus, high-quality, multi-dimensional, fine-grained feedback is essential for using LLM feedback in text summarization, with around 50K feedback pairs being a reasonable fit for preference learning.

Human and Synthetic Feedback. In Table 10, we compare the effectiveness of using FeedSum's feedback from C4 with (1) human feedback on real summaries in UniSumEval (Lee et al., 2024), and (2) synthetic feedback on synthesized summaries in SynFacEdit (Mishra et al., 2024). We train Llama3-8b-instruct using DPO, but with different feedback from the three datasets. Despite the limited size, DPO with human feedback in UniSumEval improves summary quality across all dimensions, raising the average score from 0.632 to 0.666 compared to Llama3-8b-instruct without DPO. Although the feedback in FeedSum is obtained through automated evaluation, it shows greater improvements than using 1K human feedback. Enhancements are more pronounced when increasing the size of feedback. On the other hand, the synthetic feedback in SynFacEdit decreases the quality of generated summaries after DPO, likely due to its limited size and focus on clinical summarization.

6 Conclusion

This work presents a framework for improving text summarization using LLM-generated feedback. We demonstrate that this approach is the most effective when the feedback is high-quality, multi-dimensional, and assessed at a fine-grained level. Our experiments show that DPO significantly outperforms SFT variants in utilizing such feedback. Additionally, we provide insights into the alignment trade-offs in summarization, the impact of feedback size, and the advantages of our LLM-generated feedback over existing human and synthetic alternatives. As part of our contribution, we have open-sourced both the model and dataset.

Limitations

DPO is a widely used approach for preference optimization; however, it has limitations in handling multi-dimensional feedback. A typical method involves computing a composite score by averaging the scores across all dimensions with equal weights, which may not be the optimal solution for multi-dimensional preference learning. Although we include a baseline of post-hoc parameter merging (Jang et al., 2023), recent work suggests there are better performance alternatives, such as Controllable DPO (Guo et al., 2024) and Sequential Alignment (Lou et al., 2024). We will explore the extent of performance improvement achieved by these solutions in future work.

We conducted both human and automated evaluations. However, the majority of the evaluations were automated due to the high cost associated with fine-grained, multi-dimensional manual assessments. Nevertheless, we believe that the automated evaluations provide convincing evidence, as they have demonstrated performance comparable to human evaluations (Song et al., 2024; Tang et al., 2024a; Liu et al., 2023a).

Ethics Statement

Our work primarily focuses on leveraging LLM-generated feedback on diverse text summaries, which does not pose any ethical concerns during the model training phase. For human evaluation, we followed a well-defined evaluation protocol in the literature, preventing possible ethical issues in the annotation process. Annotators were paid 50% more than the average U.S. minimum wage and received bonuses for maintaining consistent, high-quality performance.

Scientific Artifacts

The summaries used to collect LLM feedback were generated by 13 different language models. For open-source models, we used publicly available checkpoints from Huggingface, while for proprietary models, we utilized paid API services provided by OpenAI and AWS Bedrock. See Table 11 for details in Appendix.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmed Allam. 2024. Biasdpo: Mitigating bias in language models through direct preference optimization. *arXiv preprint arXiv:2407.13928*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *EMNLP*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *ACL*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *NAACL*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. QLoRA: Efficient finetuning of quantized llms. In *NeurIPS*.

Sujan Dutta, Sayantan Mahinder, Raviteja Anantha, and Bortik Bandyopadhyay. 2024. Applying rlaif for code generation with api-usage in lightweight llms. In *ACLW*.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *NAACL*.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *NAACL*.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Jiexin Wang, Huimin Chen, Bowen Sun, Ruobing Xie, Jie Zhou, Yankai Lin, et al. 2024. Controllable preference optimization: Toward controllable multi-objective alignment. *arXiv preprint arXiv:2402.19085*.

Yang Han, Yiming Wang, Rui Wang, Lu Chen, and Kai Yu. 2024. AlignSum: Data pyramid hierarchical fine-tuning for aligning with human summarization preference. In *EMNLP*.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. Meetingbank: A benchmark dataset for meeting summarization. In *ACL*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *ACL*.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2024b. A survey on human preference learning for large language models. *arXiv preprint arXiv:2406.11191*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

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.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. RLAIF: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Yuho Lee, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. 2024. UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for llms. In *EMNLP*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Xiaochen Li, Zheng-Xin Yong, and Stephen H Bach. 2024. Preference tuning for toxicity mitigation generalizes across languages. *arXiv preprint arXiv:2406.16235*.

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. G-Eval: Nlg evaluation using gpt-4 with better human alignment. In *EMNLP*.

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

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan. 2023b. On improving summarization factual consistency from natural language feedback. In *ACL*.

Xingzhou Lou, Junge Zhang, Jian Xie, Lifeng Liu, Dong Yan, and Kaiqi Huang. 2024. Spo: Multi-dimensional preference sequential alignment with implicit reward modeling. *arXiv preprint arXiv:2405.12739*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *ACL*.

Prakamya Mishra, Zonghai Yao, Parth Vashisht, Feiyun Ouyang, Beining Wang, Vidhi Dhaval Mody, and Hong Yu. 2024. SYNFAC-EDIT: Synthetic imitation edit feedback for factual alignment in clinical summarization. In *EMNLP*.

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*.

Michael Noukhovitch, Samuel Lavoie, Florian Strub, and Aaron C Courville. 2023. Language model alignment with elastic reset. In *NuerIPS*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In *NAACL*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, et al. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *ACL*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Hwanjun Song, Igor Shalyminov, Hang Su, Siffi Singh, Kaisheng Yao, and Saab Mansour. 2023. Enhancing abstractiveness of summarization models through calibrated distillation. In *EMNLP*.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using llms. In *ACL*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *NeurIPS*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. MiniCheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, et al. 2024b. TofuEval: Evaluating hallucinations of llms on topic-focused dialogue summarization. In *NAACL*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

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*.

Tempest A van Schaik and Brittany Pugh. 2024. A field guide to automatic evaluation of llm-generated summaries. In *SIGIR*.

David Wan, Koustuv Sinha, Srini Iyer, Asli Celikyilmaz, Mohit Bansal, and Ramakanth Pasunuru. 2024. ACUEVAL: Fine-grained hallucination evaluation and correction for abstractive summarization. In *ACL*.

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

Shentao Yang, Shujian Zhang, Congying Xia, Yihao Feng, Caiming Xiong, and Mingyuan Zhou. 2023. Preference-grounded token-level guidance for language model fine-tuning. *NeurIPS*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML*.

Shengqiang Zhang, Xingxing Zhang, Hangbo Bao, and Furu Wei. 2022. Attention temperature matters in abstractive summarization distillation. In *ACL*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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 *EMNLP*.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. In *NAACL*.

Model Name	Checkpoints
Bart-large	facebook/bart-large-cnn
Pegasus-large	google/pegasuscnn_dailymail
Flan-t5-large	spacemanidol/flan-t5-large-cnndm
Phi-2	microsoft/phi-2
Mistral-7b-inst	mistralai/Mistral-7B-Instruct-v0.2
Mixtral-8x7b-inst	mistralai/Mixtral-8x7B-Instruct-v0.1
Llama2-13b-chat	meta-llama/Llama-2-13b-chat-hf
Mistral-nemo	mistralai/Mistral-Nemo-Instruct-2407
Gemma2-27b-inst	google/gemma-2-27b-it
Llama3-70b	meta-llama/Meta-Llama-3-70B-Instruct
Claude-instant	claude-instant (AWS Bedrock)
GPT-3.5$_{\text{turbo}}$	gpt-3.5-turbo-0125 (OpenAI)
GPT-4$_{\text{turbo}}$	gpt-4-0125-preview (OpenAI)

Table 11: Checkpoints of the 13 summarizers. For open-source models, we use publicly available checkpoints from Huggingface, while for proprietary models, we utilize paid API services by OpenAI and AWS Bedrock.

You will receive an article along with a summary of that article.

Please evaluate the quality of summary on a Likert-scale score from 1 (bad) to 5 (perfect).

Provide your answer in JSON foramt. The answer should be a dictionary whose key is "score":
{"score": "your score"}

Source Text:
{source text}

Summary:
{summary}

JSON Output:

Table 12: Prompt to generate low-quality, single-dimensional, fine-grained feedback using C1.

A Data Creation Details

A.1 Feedback Generation

We generate LLM-based feedback across four different setups, as summarized in Table 2:

- **C1:** This setup is designed to acquire low-quality, coarse-grained, single-dimensional feedback. We perform automated evaluation using the prompt in Table 12 with Llama3-8-instruct, a lower-performing model compared to its larger counterpart, Llama3-70b-instruct. The feedback obtained is a Likert-scale overall score for the summary.

- **C2:** The prompt for this setup is identical to that of C1, but we use a nearly 10× larger LLM, Llama3-70b-instruct, to generate high-quality, single-dimensional, and coarse-grained feedback.

- **C3:** We use G-Eval (Liu et al., 2023a) with simple modification to tune it for our three key dimensions, namely faithfulness, completeness, and conciseness. We perform automated evaluation using the three prompts in Table 17. The feedback obtained is three Likert-scale scores for the three dimensions.

- **C4:** We use FineSurE (Song et al., 2024) to acquire high-quality, multi-dimensional, and fine-grained feedback at the sentence level for faithfulness; and at the key-fact level for completeness and conciseness. It performs a fact-checking task for the former and a key-fact alignment task for the latter using LLMs. The prompts for the two tasks are presented in Table 18. The feedback obtained is three percentage (%) scores for the three dimensions.

You will be provided with a transcript. Your task is to decompose the summary into a set of "key facts".
A "key fact" is a single fact written as briefly and clearly as possible, encompassing at most 2-3 entities.

Here are nine examples of key facts to illustrate the desired level of granularity:
* Kevin Carr set off on his journey from Haytor.
* Kevin Carr set off on his journey from Dartmoor.
* Kevin Carr set off on his journey in July 2013.
* Kevin Carr is less than 24 hours away from completing his trip.
* Kevin Carr ran around the world unsupported.
* Kevin Carr ran with his tent.
* Kevin Carr is set to break the previous record.
* Kevin Carr is set to break the record by 24 hours.
* The previous record was held by an Australian.

Instruction:
First, read the summary carefully. Second, decompose the summary into (at most 16) key facts.

Provide your answer in JSON format. The answer should be a dictionary with the key "key facts" containing the key facts as a list:
{"key facts": ["first key fact", "second key facts", "third key facts"]}

Summary: {summary}
JSON Output:

Table 13: Prompt to extract the list of key-facts from the reference (human) summary of original datasets.

A.2 Key-fact Extraction

The feedback from C4 requires fine-grained evaluation using key-facts to assess the completeness and conciseness scores. The key-facts are automatically extracted from the reference (human) summary of each source dataset, as suggested by Song et al. (2024). Thus, we obtain the list of key-facts for 15.4K documents in FeedSum: 14K for training set and the remaining 1.4K for testing set. The prompt used for automated key-fact extraction is detailed in Table 13.

Dataset	Type	Document Length	Domain	# of Document	Document Word count (Min – Max)	Summary Word count (Min – Max)	Key-fact Count (Min – Max)
CNNDM	Non-Dialogue	Short	News	21194	675.8 (46–1919)	48.8 (10–162)	6.3 (1–16)
WikiHow			Lifestyle	21980	72.9 (10–680)	6.56 (1–51)	1.4 (1–16)
GovReport		Long	Report	8066	3573.0 (141–5873)	439.4 (29–1002)	14.7 (4–21)
PubMed			Medical	17843	2491.0 (10–6384)	210.6 (49–402)	12.0 (3–26)
DialogSum	Dialogue	Short	Daily Life	21957	122.4 (33–727)	22.6 (5–101)	3.7 (1–16)
MediaSum		Long	Interview	18927	1373.7 (80–5111)	14.3 (5–97)	2.6 (1–14)
MeetingBank			Meeting	15421	1283.1 (96–5803)	56.4 (14-184)	6.8 (1–20)

Table 14: Statistics of the FeedSum training set, detailing the average word count of input documents, reference summaries, and key-facts, with respective min-max ranges in parentheses (reference summaries refer to the human-written summaries in the original datasets). Documents with over 1K words are considered "long".

A.3 Dataset Statistic

We present a comprehensive statistical analysis of the FeedSum datasets, which consist of 125,388 <document, summary, feedback> triplets for each configuration outlined in Table 2. Detailed statistics of FeedSum are provided in Table 14.

B Training Detail

B.1 Training Configuration

For preference learning, we investigate two possible solutions of supervised fine-tuning (SFT) and direct preference optimization (DPO). The details of each configuration are detailed below:

Supervised Fine-tuning (SFT). We fine-tune Llama3-8b-instruct using QLoRA (Dettmers et al., 2024) and DeepSpeed (Stage-2) (Rasley et al., 2020) on four NVIDIA H100 GPUs. The model is trained for 3,000 steps with AdamW as the optimizer, using a batch size of 32, an initial learning rate of 1e-4, and a weight decay of 0.05. Regardless of how to select the reference summary, we apply the same configuration for all SFT strategies, namely SFT-{human, best, faith, comp, conc} in Table 8. The input (user prompt) and output (assistant prompt) for Llama3-8b-instruct are configured similarly to the example for DPO in Table 19. The difference is that SFT only passes the input along with a single output, selected based on a predefined criterion, e.g., the summary with the highest composite score.

Direct Preference Optimization (DPO). We train Llama3-8b/70b-instruct using DPO (Rafailov et al., 2023). Since the model has completed the instruction-tuning process, we proceed directly to optimize it using DPO. Like SFT, we apply QLoRA and DeepSpeed (Stage-2) to train the model on four NVIDIA H100 GPUs. The model is trained for 6,000 steps with AdamW as the optimizer, using a batch size of 32, an initial learning rate of 5e-5, and a weight decay of 0.05. We apply the same setup for all configurations, namely C1 – C4 in Table 2. The input (user prompt) and output (assistant prompt) for Llama3-8b-instruct are configured in the example of Table 19.

For ablation studies for feedback type and size in Section 5.4, we adjust the number of steps due to the different number of human or LLM-generated feedback. We reduce the number of steps to 4,000 when the number of feedback pairs exceeds 40,000; otherwise, we reduce it to 3,000. Other configurations remain the same.

B.2 Input and Output Format

Tables 19 presents an example of the input and its corresponding chosen and rejected outputs to train Llama3-8b-instruct using DPO. We follow the same prompt style of Llama3 for instruction tuning. In this example, the chosen summary was generated by GPT-4-turbo, achieving scores of 100% for faithfulness, 60% for completeness, and 100% for conciseness. On the other hand, the rejected summary was generated by Mistral-7b-instruct, achieving scores of 25% for faithfulness, 60% for completeness, and 50% for conciseness.

The auto-evaluation results of the example by FineSurE (Song et al., 2024) are provided in Table 20. The final percentage (%) scores can be computed by calculating the proportion of factually correct sentences for faithfulness, that of included given key-facts for completeness, and that of summary sentences related to the key-facts. The detailed equation is provided in Appendix C.1.

For SFT, the input and its corresponding response are almost similar to those of DPO. But,

there is no distinction between chosen and rejected summaries. We select reference summaries using five different criteria: SFT-human, SFT-best, SFT-faith, SFT-comp, and SFT-cons, as detailed in Section 3.3. The selected reference summaries are provided to train Llama3 using teacher-forcing.

C Automated and Human Evaluation

C.1 Metric for Summary Quality

We utilize three dimensions of metrics, namely faithfulness, completeness, and conciseness, along with one that estimates the abstractiveness of the summary, in line with recent literature (Song et al., 2024; Lee et al., 2024).

Faithfulness Score. Faithfulness score is formulated by aggregating sentence-level fact check results. Let $S = \{s_1, \ldots, s_N\}$ is the summary passage which consists of N sentences, where s_i denotes the i-th sentence in the summary passage. Let $S_{\text{fact}} \subseteq S$ represent the subset of sentences verified as "factually correct." The faithfulness percentage score of S, with respect to the document D, is defined as:

$$\text{Faithful}(D, S) = |S_{\text{fact}}|/|S|. \quad (1)$$

This metric measures the proportion of factually correct sentences in the summary relative to the total number of sentences in the summary.

Completeness and Conciseness Score. Let $K = \{k_1, \ldots, k_M\}$ be the collection of key-facts, where M indicates the total number of these facts. Utilizing the results from the alignment of key-facts, we can establish a bipartite graph $M = (K, S, E)$, with set of edges $E = \{(k, s) : k \rightarrow s \,|\, k \in K \wedge s \in S\}$. Here, the notation $k \rightarrow s$ signifies that the key-fact k is identified as being included in the summary sentence s. The completeness and conciseness scores for summary S are computed as percentage scores, defined as follows:

$$\text{Complete}(K, S) = |\{k \,|\, (k, s) \in E\}|/|K|, \quad (2)$$

$$\text{Concise}(K, S) = |\{s \,|\, (k, s) \in E\}|/|S|. \quad (3)$$

In this context, the operator $|\cdot|$ denotes the cardinality of a set. Completeness score indicates how well the key-facts are incorporated into the summary. Furthermore, the conciseness score evaluates how effectively the summary condenses and includes the key-facts.

Composite Score. To determine the chosen and rejected summaries in cases of multi-dimensional feedback, we use the average of the three percentage scores – faithfulness, completeness, and conciseness – to calculate a composite score.

Abstractiveness Score. In Section 5.3, we additionally report the abstractiveness score of the summary, which refers to the extent to which a summary generates novel sentences or phrases, leading to a more coherent summary. The abstractiveness score is measured by calculating the ratio of novel n-grams present in the summary that does not appear in the original input text (Liu and Lapata, 2019; Song et al., 2023). Let $n\text{-gram}_{\text{shared}}$ represent the set of n-grams that are shared between the summary and the document, while $n\text{-gram}_{\text{summary}}$ denotes the total set of n-grams included in the summary. Then, the ratio of novel n-grams N_n is defined as:

$$N_n = 1 - |n\text{-gram}_{\text{shared}}|/|n\text{-gram}_{\text{summary}}|. \quad (4)$$

The final abstractiveness score for a summary S is computed as the average of the novel 1/3/5-gram ratios, as follows:

$$\text{Abstractive}(D, S) = (N_1 + N_3 + N_5)/3. \quad (5)$$

C.2 Metric for Feedback Quality

In table 3, we use the same settings as in recent studies (Liu et al. 2023b, Song et al. 2024) to evaluate the summary feedback quality and align it with human judgment. Specifically, there are two levels for evaluating the alignment (correlation) of generated summary feedback with human feedback. The greater the alignment, the higher the quality of the generated feedback.

Summary-level Correlation. We can check the alignment between the generated and human feedback at the summary level. To calculate the summary-level correlation, we define F_{actual} and F_{pred} as the percentage scores of the ground truth and the predicted summaries, respectively. Let $D = \{D_1, \ldots, D_k\}$ represent the set of input documents, and $S = \{S_1, \ldots, S_k\}$ represent the corresponding summaries for these documents. The summary-level correlation is computed as:

$$\text{Spearman}([F_{actual}(D_1, S_1), \ldots, F_{actual}(D_k, S_k)], [F_{pred}(D_1, S_1), \ldots, F_{pred}(D_k, S_k)]). \quad (6)$$

Method	Backbone	ROUGE-1	ROUGE-2	ROUGE-L	BERT-F1	BERT-P	BERT-R
wo. RL	Llame3-8b-inst.	0.453	0.172	0.231	0.854	0.841	0.867
wo. RL	Llame3-70b-inst.	0.450	0.183	0.213	0.854	0.843	0.867
DPO-C1	Llame3-8b-inst.	0.435	0.182	0.240	0.853	0.840	0.866
DPO-C2	Llame3-8b-inst.	**0.480**	0.187	**0.257**	0.856	0.846	0.867
DPO-C3	Llame3-8b-inst.	0.419	0.168	0.231	0.856	**0.847**	0.866
DPO-C4	Llame3-8b-inst.	0.474	**0.188**	0.234	**0.857**	0.846	**0.869**

Table 15: Results using two conventional automated metrics on six summarizers with and without DPO: ROUGE-{1, 2, L} and BERTScore-{F1, Precision, Recall}. The best scores are marked in bold.

We employ the Spearman correlation as our correlation measure. Ultimately, the summary-level correlation reflects the alignment between human-assessed feedback and LLM-generated feedback for the identical document.

System-level Correlation. The system-level evaluation assesses the alignment of performance rankings across summarization systems (summarizers) as determined by both our LLM feedback scores and human feedback scores. To calculate the system-level rank correlation, we consider $\mathbf{F}_m = \{F_m(D_1, S_1), \ldots, F_m(D_M, S_M)\}$ as the set of percentage scores derived from M document-summary pairs generated by the summarization model m. Next, we construct a list of the average percentage scores for all T summarization models, denoted as $\left[\bar{\mathbf{F}}_{m_1}, \ldots, \bar{\mathbf{F}}_{m_T}\right]$ where, $\bar{\mathbf{F}}_m = \frac{1}{M}\sum_{i=1}^{M} F_m(D_i, S_i)$. Applying the rank function to this list, we derive the ranking list $\text{rank}\left([\bar{\mathbf{F}}_{m_1}, \ldots, \bar{\mathbf{F}}_{m_T}]\right) = [\text{rank}_{m_1}, \ldots, \text{rank}_{m_T}]$, where rank_m represents the rank of model m:

$$\text{rank}(\bar{\mathbf{F}}_{m_i}) = \sum_{j=1}^{T} \mathbf{1}(\bar{\mathbf{F}}_{m_j} \leq \bar{\mathbf{F}}_{m_i}). \quad (7)$$

We derive the rank list $[\text{rank}_{m_1}, \ldots, \text{rank}_{m_T}]$ based on LLM feedback scores, as well as the rank list $\left[\text{rank}^*_{m_1}, \ldots, \text{rank}^*_{m_T}\right]$ based on human feedback scores. Then, the system-level correlation is computed as:

$$\text{Spearman}([\text{rank}_{m_1}, \ldots, \text{rank}_{m_T}], \left[\text{rank}^*_{m_1}, \ldots, \text{rank}^*_{m_T}\right]). \quad (8)$$

The system-level rank correlation evaluates the degree of agreement between the rankings generated from LLM feedback scores and human-provided feedback scores across different summarization systems.

C.3 Human Evaluation Details

Fine-Grained Annotation Tasks We conduct two human annotation tasks: (1) fact verification and (2) key-fact alignment. The format of the two annotation tasks draw on the annotation protocol suggested by Lee et al. (2024). In fact verification, annotators assign a binary label (Yes/No) to indicate whether a sentence contains factual errors. For key-fact alignment, annotators evaluate whether summary sentences contain key-fact of their source text. We use the human-verified key-facts from the existing dataset created by Lee et al. (2024).

For the two annotation tasks, we compute percentage scores for three summary-level evaluation dimensions: (1) faithfulness, the proportion of factually accurate sentences; (2) completeness, the percentage of key-facts covered by the summary; and (3) conciseness, the proportion of sentences relevant to the key-facts. The detailed formulation can be found in Appendix C.1.

Annotator Qualifications and Costs We used Amazon Mechanical Turk (MTurk) annotators with an approval rating above 95% and at least 1,000 accepted HITs. A detailed qualification test of English comprehension questions, simulating the actual annotation tasks, was required. We only recruited annotators who received the perfect score on the test and resided in AU, CA, NZ, GB, or the US. The total cost of human annotation exceeded $750 for 420 input text-summary pairs, with payments above the U.S. minimum wage.

D Results with Conventional Metric

We conduct automated evaluation using two popular conventional metrics, namely ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019). Although it has been recently recognized that these metrics do not align well with human evaluations of text summaries (Pagnoni et al., 2021; Song et al., 2024), these scores can still serve as auxiliary metrics to assess word overlap (using ROUGE) and semantic relevance (using BERTScore) with the given reference summaries.

Table 15 presents the ROUGE and BERT scores on the test set of FeedSum. We use the human-

Backbone	Config.	Quality	Dimension	Granularity	Faith.	Comp.	Conc.	Avg.
Gemma-2b-inst.	wo. RL				0.558	0.361	0.422	0.447
Gemma-2b-inst.	DPO-C1	Low	Single	Coarse-grained	0.507	0.373	0.463	0.448
	DPO-C2	High	Single	Coarse-grained	0.556	0.383	0.498	0.479
	DPO-C3	High	Multiple	Coarse-grained	0.588	0.384	0.481	0.484
	DPO-C4	High	Multiple	Fine-grained	**0.613**	**0.396**	**0.533**	**0.514**

Table 16: Automated evaluation results of five summarizers on faithfulness, completeness, and conciseness. "w.o RL" refers to the models without preference optimization. "DPO-{C1,C2,C3,C4}" denote models incorporating preference optimization, utilizing feedback generated in C1–C4 of Table 2. The best scores are marked in bold.

written summaries from the six source datasets as reference summaries to compute the respective scores. All the summarizers demonstrate consistently high ROUGE and BERT scores, both with and without DPO. While there is no significant difference between them, DPO-C4 achieves the highest scores in three out of six score categories.

E Prompt for Summary Generation

We use two distinct prompts for summary generation: (1) FeedSum Benchmark, where we generate summaries of varying quality using 13 different language models[4] as summarizers, and (2) Evaluation of summarizers (i.e., Llama3 variants) after SFT or DPO.

For the former, we use a simple prompt:
INSTRUCTION: SUMMARIZE THE TEXT.
PROVIDE YOUR ANSWER IN JSON FORMAT. THE ANSWER SHOULD BE A DICTIONARY WITH THE KEY "SUMMARY" CONTAINING A GENERATED SUMMARY AS A STRING: {"SUMMARY": "YOUR SUMMARY"}.

For the latter, we use the exact same prompt to generate summaries across all Llama3 variants, identical to the input prompt shown in Table 19.

F Results with Gemma-2b-instruct

We conduct an additional experiment to evaluate the improvements from preference learning with LLM-generated feedback, using a different LLM. Specifically, we choose Gemma-2b-instruct, as this smaller model demonstrates the impact of our framework, even when compared to significantly larger models like Llama-8/70b-instruct.

Table 16 presents the automated evaluation results across three fine-grained dimensions of summary quality, comparing their percentage scores before and after applying DPO with feedback from C4. LLM-generated feedback leads to significant improvements across all dimensions. However, these improvements are smaller than those observed in the Llama3 family (Table 5), indicating that larger LLMs benefit more from preference optimization with LLM-generated feedback in the context of text summarization.

Lastly, the significant performance gap between Gemma and Llama3 is primarily due to Gemma's inability to generate summaries for longer documents. When we test it, Gemma produces a weird response, e.g., THE AIM WAS A PRIORI KNOWLEDGE ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT ABOUT (...). This suggests that 2-billion-parameter models may not be suitable for long document summarization.

G Summary Example

Table 21 presents examples of summaries generated by six different approaches: Llama3-8b/70b-instruct without DPO, and four variants of Llama3-8b-instruct after applying DPO.

The summary of DPO-C4 can be considered the best for the following reasons:

- **Core Focus:** The summary accurately captures the main theme of the conversation, which revolves around the Thanksgiving dinner arrangements. It highlights how the two people confirm plans, discuss what to bring, and finalize the decision for Person2 to bring wine instead of pie. This maintains the core context.
- **Inclusion of Key-facts:** The summary covers the important details of the conversation, including Person2's initial offer to bring dessert (pumpkin pie) and the shift to bringing wine due to another family member handling dessert. Other summaries tend to overlook or simplify this progression, while DPC-C4 fully captures the interaction's key events.
- **Clarity and Conciseness:** The summary is structured in a straightforward, concise manner, effectively summarizing the conversation without un-

[4]Non-LLMs, such as Bart-large and Pegasus-large, do not require any prompt as they directly generate summaries from input documents. This is because they were trained via supervised fine-tuning on summarization datasets like CNNDM.

necessary details. It presents the flow and outcome of the discussion clearly, making it easy for readers to understand. The logical order of events is maintained, ensuring a smooth narrative.

• **Accurate Role Depiction:** The summary clearly identifies Person1 as the host and Paul (Person2) as the guest, which helps clarify their relationship and the nature of the conversation. This distinction is more explicit in DPC-C4 compared to other summaries, which might leave these roles more ambiguous.

In conclusion, DPC-C4 is the best summary because it captures the essential points of the conversation with clarity and completeness, while maintaining a concise and well-structured form. It ensures that all significant details are included without overwhelming the reader.

Faithfulness	You will be given an article. You will then be given one summary written for this article. Your task is to rate the summary on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Consistency (1-5) - the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document. Annotators were also asked to penalize summaries that contained hallucinated facts. Evaluation Steps: 1. Read the news article carefully and identify the main facts and details it presents. 2. Read the summary and compare it to the article. Check if the summary contains any factual errors that are not supported by the article. 3. Assign a score for consistency based on the Evaluation Criteria. Example: Source Text: {source text} Summary: {summary} Evaluation Form (scores ONLY): - Consistency:
Completeness	You will be given an article. You will then be given one summary written for this article. Your task is to rate the summary on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Completeness (1-5) - the degree to which the summary includes all key information present in the source document. A complete summary accurately captures the main points, ideas, and relevant details without omitting crucial elements. Evaluation Steps: 1. Read the news article carefully and identify the main points, key information, and relevant details. 2. Read the summary and compare it to the article. Check if the summary captures all essential facts, main ideas, and pertinent details presented in the original article. 3. Assign a score for completeness based on the Evaluation Criteria. Example: Source Text: {source text} Summary: {summary} Evaluation Form (scores ONLY): - Completeness:
Conciseness	You will be given an article. You will then be given one summary written for this article. Your task is to rate the summary on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed. Evaluation Criteria: Conciseness (1-5) - the extent to which the summary presents information succinctly and without unnecessary elaboration. A concise summary effectively conveys the essential content of the source document using clear and concise language, avoiding redundant or superfluous information. Evaluation Steps: 1. Read the news article carefully and identify the main points, key information, and relevant details. 2. Read the summary and compare it to the article. Check if the summary effectively conveys the essential content of the document in a concise manner, without unnecessary elaboration or redundancy. 3. Assign a score for conciseness based on the Evaluation Criteria. Example: Source Text: {source text} Summary: {summary} Evaluation Form (scores ONLY): - Conciseness:

Table 17: Prompt of the modified G-Eval for faithfulness, completeness, and conciseness.

Fact Check	You will receive a document followed by a corresponding summary. Your task is to assess the factuality of each summary sentence across nine categories: * no error: the statement aligns explicitly with the content of the document and is factually consistent with it. * out-of-context error: the statement contains information not present in the document. * entity error: the primary arguments (or their attributes) of the predicate are wrong. * predicate error: the predicate in the summary statement is inconsistent with the document. * circumstantial error: the additional information (like location or time) specifying the circumstance around a predicate is wrong. * grammatical error: the grammar of the sentence is so wrong that it becomes meaningless. * coreference error: a pronoun or reference with wrong or non-existing antecedent. * linking error: error in how multiple statements are linked together in the discourse (for example temporal ordering or causal link). * other error: the statement contains any factuality error which is not defined here. Instruction: First, compare each summary sentence with the document. Second, provide a single sentence explaining which factuality error the sentence has. Third, answer the classified error category for each sentence in the summary. Provide your answer in JSON format. The answer should be a list of dictionaries whose keys are "sentence", "reason", and "category": ["sentence": "first sentence", "reason": "your reason", "category": "no error", "sentence": "second sentence", "reason": "your reason", "category": "out-of-context error", "sentence": "third sentence", "reason": "your reason", "category": "entity error",] Document: {document} Summary with {# of sentences} sentences: {sentences} JSON Output:
Key-fact Alignment	You will receive a summary and a set of key facts for the same document. Your task is to assess if each key fact is inferred from the summary. Instruction: First, compare each key fact with the summary. Second, check if the key fact is inferred from the summary and then response "Yes" or "No" for each key fact. If "Yes", specify the line number(s) of the summary sentence(s) relevant to each key fact. Provide your answer in JSON format. The answer should be a list of dictionaries whose keys are "key fact", "response", and "line number": ["key fact": "first key fact", "response": "Yes", "line number": [1], "key fact": "second key fact", "response": "No", "line number": [], "key fact": "third key fact", "response": "Yes", "line number": [1, 2, 3]] Summary: {summary} {# of key facts} key facts: {key facts} JSON Output:

Table 18: Prompt of the FineSurE for fact checking and key-fact alignment tasks.

Input	<|begin_of_text|><|start_header_id|>user<|end_header_id|> Below is an instruction that describes a task. Write a response that appropriately completes the request. ###Instruction: Please summarize the input document. ###Input: MICHELE NORRIS, Host: And as NPR's Jackie Northam reports, Zardari isn't getting much relief while he's here.</s>JACKIE NORTHAM: Brian Katulis, a senior fellow at the Center for American Progress, says the attack sent a clear signal to the Obama administration that despite efforts to curb militancy, there's a serious undercurrent of extremism that has grown and metastasized in Pakistan.</s>BRIAN KATULIS: The concern of infiltration of those extremist voices in Pakistani's security services, a country that has anywhere from 70 to a hundred nuclear weapons, is the thing that keeps many people in Washington up at night.</s>JACKIE NORTHAM: Security and counterterrorism efforts are among the issues discussed today by Presidents Zardari and Obama before the Holbrooke service. But Katulis says Zardari isn't the best man for that discussion because he doesn't hold much sway.</s>BRIAN KATULIS: The meeting between President Obama and Zardari is an example of head-of-state diplomacy at its most complicated. Zardari, as the head of civilian government, really doesn't have as much power over those security issues as do Ashfaq Kayani, the head of the Pakistani military, or Ahmad Shuja Pasha who's the head of the Inter-Services Intelligence.</s>JACKIE NORTHAM: Jamie Metzl, the executive vice president of the Asia Society, says Pakistan needs to do better to justify U.S. support.</s>JAMIE METZL: We've spent $20 billion in Pakistan since 9/11, huge amounts of military aid; lately, more civilian-focused aid, and the situation in Pakistan seems to have gone from bad to worse.</s>JACKIE NORTHAM: Kamran Bokhari, with the intelligence firm STRATFOR, says there are two schools of thought in Washington over how to deal with Pakistan. One is that Pakistan is playing a double game with Washington.</s>KAMRAN BOKHARI: This view says we need to be able to sustain the pressure on Pakistan, they can definitely do more, they're just not doing it. On the other hand, there are those who say Pakistan is already quite weakened. So if we demand more from the Pakistanis, what that means is that there is a good chance that it could further undermine stability within Pakistan.</s>JACKIE NORTHAM: Jackie Northam, NPR News, Washington. ###Response:<|eot_id|>
Output (Chosen)	<|start_header_id|>assistant<|end_header_id|> NPR's Jackie Northam reports on concerns in Washington regarding extremism in Pakistan, highlighting the challenges faced by President Zardari in addressing security and counterterrorism with President Obama. Experts like Brian Katulis and Jamie Metzl criticize Pakistan's handling of militancy and question the effectiveness of U.S. aid, given the country's deteriorating situation. Kamran Bokhari of STRATFOR outlines the debate in Washington on how to approach Pakistan, balancing the need for pressure against the risk of destabilizing the country further.<|eot_id|>
Output (Rejected)	<|start_header_id|>assistant<|end_header_id|> Pakistani President Asif Ali Zardari arrived in Washington for a meeting with President Obama, but has faced mounting pressure at home following the attack on the U.S. Embassy in Islamabad. The attack has raised concerns that extremist groups may have infiltrated Pakistan's security services, which oversee the country's nuclear arsenal. Zardari is seen as having little influence over security issues, with the military and intelligence agencies wielding more power. The meeting between the two presidents was focused on security and counterterrorism efforts in Pakistan, but experts warn that Pakistan's instability could continue to pose a danger to U.S. interests.<|eot_id|>

Table 19: An example of an input (user prompt) and its corresponding chosen and rejected outputs (assistant prompt) used to train Llama3-8b-instruct using DPO.

Summarizer	FineSurE's Results
GPT-4-turbo	**Task 1. Fact Checking** [{"sentence": "NPR's Jackie Northam reports on concerns in Washington regarding extremism in Pakistan, highlighting the challenges faced by President Zardari in addressing security and counterterrorism with President Obama.", "reason": "The sentence accurately summarizes the content of the document.", "category": "no error"}, {"sentence": "Experts like Brian Katulis and Jamie Metzl criticize Pakistan's handling of militancy and question the effectiveness of U.S. aid, given the country's deteriorating situation.", "reason": "The sentence accurately summarizes the opinions of Brian Katulis and Jamie Metzl in the document.", "category": "no error"}, {"sentence": "Kamran Bokhari of STRATFOR outlines the debate in Washington on how to approach Pakistan, balancing the need for pressure against the risk of destabilizing the country further.", "reason": "The sentence accurately summarizes Kamran Bokhari's views on the debate in Washington.", "category": "no error"}] **Faithfulness Score = 3/3 = 100%** **Task 2. Key-fact Alignment** [{"key fact": "Pakistani President Asif Ali Zardari is visiting Washington this week", "response": "Yes", "line number": [1]}, {"key fact": "Asif Ali Zardari is attending a memorial for Richard Holbrooke", "response": "No", "line number": []}, {"key fact": "Officials describe Zardari's visit as private", "response": "No", "line number": []}, {"key fact": "Pakistan is currently experiencing political turmoil", "response": "Yes", "line number": [2, 3]}, {"key fact": "The U.S. is pressuring Pakistan to curb terrorism", "response": "Yes", "line number": [1, 3]}] **Completeness Score = 3/5 = 60%** **Conciseness Score = 3/3 = 100%**
Mistral-7b-inst	**Task 1. Fact Checking** [{"sentence": "Pakistani President Asif Ali Zardari arrived in Washington for a meeting with President Obama, but has faced mounting pressure at home following the attack on the U.S. Embassy in Islamabad.", "reason": "The document does not mention the attack on the U.S. Embassy in Islamabad.", "category": "out-of-context error"}, {"sentence": "The attack has raised concerns that extremist groups may have infiltrated Pakistan's security services, which oversee the country's nuclear arsenal.", "reason": "The document mentions concerns about infiltration of extremist voices in Pakistani's security services, but not specifically the attack raising these concerns.", "category": "out-of-context error"}, {"sentence": "Zardari is seen as having little influence over security issues, with the military and intelligence agencies wielding more power.", "reason": "The document explicitly states that Zardari doesn't hold much sway over security issues, and that the head of the Pakistani military and the head of the Inter-Services Intelligence have more power.", "category": "no error"}, {"sentence": "The meeting between the two presidents was focused on security and counterterrorism efforts in Pakistan, but experts warn that Pakistan's instability could continue to pose a danger to U.S. interests.", "reason": "The document mentions that security and counterterrorism efforts were discussed, but does not mention experts warning about Pakistan's instability posing a danger to U.S. interests.", "category": "out-of-context error"}] **Faithfulness Score = 1/4 = 25%** **Task 2. Key-fact Alignment** [{"key fact": "Pakistani President Asif Ali Zardari is visiting Washington this week", "response": "Yes", "line number": [1]}, {"key fact": "Asif Ali Zardari is attending a memorial for Richard Holbrooke", "response": "No", "line number": []}, {"key fact": "Officials describe Zardari's visit as private", "response": "No", "line number": []}, {"key fact": "Pakistan is currently experiencing political turmoil", "response": "Yes", "line number": [1]}, {"key fact": "The U.S. is pressuring Pakistan to curb terrorism", "response": "Yes", "line number": [4]}] **Completeness Score = 3/5 = 60%** **Conciseness Score = 2/4 = 50%**

Table 20: Results by FineSurE (Song et al., 2024) on the "chosen" and "rejected" summaries in Table 19.

Document	#Person1#: Hey, Paul, you're still having Thanksgiving dinner at my house on Thursday, right? #Person2#: Yeah, thanks again for the invitation. I was worried I'd have to spend it alone after my parents announced they'd be in Finland. #Person1#: I think it's strange that you're not going with them. Is it because of the cold weather in Northern Europe? #Person2#: Not exactly. I have a lot of work to do next week. #Person1#: Well, I'm glad you won't be alone on a holiday. It'll be great to have you celebrate with us. The more the merrier! #Person2#: Exactly, how many people will be there? #Person1#: Let's see, my parents, my grandpa, my sister, my brother-in-law, my niece and my two cousins, plus us. So, ten people altogether. #Person2#: Should I bring anything? #Person1#: Nope, we've got it covered. #Person2#: Come on, I hate to show up empty-handed. You are being so generous to have me. I know! I'll bring pie. I'll make a great pumpkin pie. #Person1#: Sounds delicious, but my sister has already volunteered for dessert. Look, don't worry about it. There will be tons of food. Turkey, stuffing, green beans, sweet potatoes. #Person2#: Wow, I'm getting hungry just thinking about it. How about this, I'll bring a nice bottle of wine. You can never have too much of that. #Person1#: Well, isn't that the truth? Sounds good to me.
Llama3-8b-inst. (wo. RL)	Person1 invites Person2 to join them for Thanksgiving dinner at their house. Person2 accepts the invitation, citing that they were worried they would have to spend the holiday alone after their parents announced they would be in Finland. Person1 reassures Person2 that they are welcome and that the more the merrier. Person2 asks how many people will be attending and is told that there will be 10 people, including their family and cousins. Person2 offers to bring a gift, but Person1 declines, saying that they have everything covered. Person2 persists and offers to bring a bottle of wine or a pie, but Person1 suggests that their sister has already volunteered to bring dessert.
Llama3-70b-inst. (wo. RL)	Person 1 invites Person 2 to their house for Thanksgiving dinner on Thursday. Person 2 is grateful for the invitation and explains they can't go to Finland with their parents due to work commitments. Person 1 is happy to have them celebrate with their family, which will consist of 10 people. When Person 2 offers to bring something, Person 1 declines, but eventually agrees to let them bring a bottle of wine.
Llama3-8b-inst. (w. DPO-C1)	Paul and Person2 confirm their plans for Thanksgiving dinner at Person1's house, discussing the guest list and what each person will bring. Person2 offers to bring a pumpkin pie, but Person1 declines, mentioning that someone else is already providing dessert. Person2 then offers to bring a bottle of wine instead, which Person1 accepts.
Llama3-8b-inst. (w. DPO-C2)	Person1 and Person2 are confirming their plans for Thanksgiving dinner at Person1's house on Thursday. Person2 is grateful for the invitation and offers to bring a contribution, initially suggesting pie, but then changes to a bottle of wine instead, since Person1's sister is already handling dessert. Person1 is appreciative of the offer, but insists that they have everything covered and that the wine sounds like a great addition.
Llama3-8b-inst. (w. DPO-C3)	Person1 invites Person2 to join their family's Thanksgiving dinner, which Person2 had initially worried they might spend alone due to their parents traveling to Finland. Person2 expresses gratitude and offers to bring something, but Person1 assures them that they have everything covered. Person2 decides to bring a bottle of wine instead, which is accepted with enthusiasm.
Llama3-8b-inst. (w. DPO-C4)	The conversation is about Thanksgiving dinner arrangements, with Person2 (Paul) confirming his attendance at Person1's (host's) house on Thursday, discussing what to bring and what's already been taken care of, and finalizing the offer to bring wine instead of pie since someone else is handling desserts.

Table 21: Example of summaries for an identical dialogue from DialogSum.