Factual Dialogue Summarization via Learning from Large Language Models

Anonymous ACL submission

Abstract

Factual consistency is an important quality in dialogue summarization. Large language model (LLM)-based automatic text summarization models generate more factually consistent summaries compared to those by smaller pretrained language models, but they face deployment challenges in real-world applications due to privacy or resource constraints. In this paper, we investigate the use of symbolic knowledge distillation to improve the factual consistency of smaller pretrained models for dialogue summarization. We employ zero-shot learning to extract symbolic knowledge from LLMs, generating both factually consistent (positive) and inconsistent (negative) summaries. We then apply two contrastive learning objectives on these summaries to enhance smaller summarization models. Experiments with BART, PEGASUS, and Flan-T5 indicate that our approach surpasses strong baselines that rely on complex data augmentation strategies. Our approach achieves better factual consistency while maintaining coherence, fluency, and relevance, as confirmed by various automatic evaluation metrics. We also provide access to the data and code to facilitate future research [1].

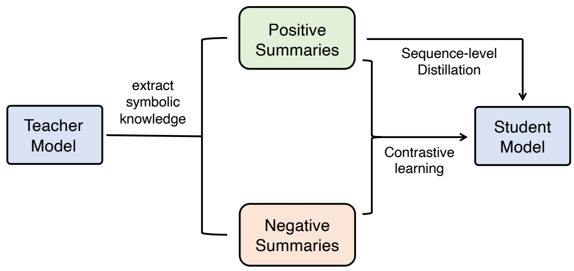

Figure 1: An overview of our framework to leverage symbolic knowledge distillation to improve the factual consistency for smaller (student) models in dialogue summarization.

1 Introduction

Automatic text summarization aims to create a concise summary of a source document that keeps all the essential points. Although current models are capable of generating fluent and coherent summaries, one main issue is factual inconsistency, where generated summaries are found to contain facts that are absent from or contradict the source (Maynez et al., 2020; Huang et al., 2021). To tackle this, a number of methods have been proposed, including explicit fact modeling (Zhu et al., 2021; Huang et al., 2020), post-editing (Lee et al., 2022; Balachandran et al., 2022; Chen et al., 2021a) and contrastive learning (Wan and Bansal, 2022a; Cao and Wang, 2021; Liu et al., 2021). Contrastive learning-based methods, in particular, offer a straightforward solution without requiring any modification to the model architecture, but their performance hinges on careful and often rule-based construction of negative samples (Cao and Wang, 2021; Liu et al., 2021; Wan and Bansal, 2022a).

The rise of large language models (LLMs) changed the landscape of NLP, and they exhibit emergent capabilities (Wei et al., 2022) such as in-context learning (Brown et al., 2020; Min et al., 2022) and instruction following (Ouyang et al., 2022). We have seen zero- or few-shot prompting with LLMs achieving strong performance on various NLP tasks (Wei et al., 2021; Ye et al., 2021) including summarization (Zhang et al., 2023), showing better coherence, relevance and factual consistency than human-written reference summaries.

Although impressive, LLMs are not always deployable in real-world applications due to substantial computational resources (Strubell et al., 2019) or privacy concerns (as many state-of-the-art LLMs are closed source and can only be accessed via APIs). Thus, it is important to construct more cost-efficient and compact models with similar summa-

[1] https://anonymous.4open.science/r/symbolic_distill_contrastive_summ-73D7/README.md

rization capabilities. To this end, knowledge distillation (Hinton et al., 2015) — a technique that can transfer the knowledge from a large *teacher model* to a small *student model* — has been explored (Sun et al., 2020; Aguilar et al., 2020). Symbolic knowledge distillation (West et al., 2022), a special form of knowledge distillation, extracts symbolic knowledge (e.g., textual information) from the teacher model and uses such knowledge as training signal for the student model. This method is especially useful when working with blackbox teacher models where we do not have access to their output probability distribution (which is the case for closed source LLMs such as ChatGPT).

In this paper, we explore symbolic knowledge distillation to improve the factual consistency of (smaller) pretrained models in dialogue summarization. Concretely, we extract symbolic knowledge from an LLM teacher (*gpt-3.5 turbo*) in the format of **positive summaries** and **negative summaries**. Positive summaries are factually consistent with the source article (i.e., a dialogue) while negative summaries are not. We experiment with various strategies to incorporate these summaries and train the student model, including sequence-level knowledge distillation (Kim and Rush, 2016) and two contrastive learning-based methods. Our experiments cover three widely used pretrained models: BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020), and Flan-T5 (Chung et al., 2024) on two popular dialogue summarization datasets: SAMSum (Gliwa et al., 2019a) and DialogSum (Chen et al., 2021b).

To summarize, our contributions are as follows:

- We propose to improve the factual consistency of (small) dialogue summarization models via symbolic knowledge distillation from LLMs.
- We experiment with LLMs to generate not only factually consistent summaries but also inconsistent ones, and we incorporate such summaries to train small dialogue summarization models with two contrastive objectives.
- We discovered that: (1) symbolic knowledge distillation enables us to create smaller dialogue summarization models that surpass strong baselines; and (2) the top-performing student model achieves comparable or even better factual consistency compared to human-written references without compromising other quality dimensions such as fluency or coherence.

2 Related Work

2.1 Evaluating and Enhancing Factual Consistency

We summarize two areas of factuality research: *evaluation* and *enhancement*.

Automatic evaluation metrics are generally constructed on question-answering systems (Fabbri et al., 2022; Scialom et al., 2021; Durmus et al., 2020; Manakul et al., 2023) or textual entailment models (Kryscinski et al., 2020; Goyal and Durrett, 2020; Laban et al., 2022; Zhang et al., 2024). More recent methods leverage the capability of LLMs to follow zero-shot and few-shot instructions (Fu et al., 2023; Min et al., 2023; Liu et al., 2023b). Another line of work aims at developing metrics that can detect the factual consistency between text pairs in different tasks (Deng et al., 2021; Zha et al., 2023a), such as a knowledge-grounded dialogue.

Methods to enhance the factual consistency of summarization models mainly fall into the following categories: explicit modeling of the facts in source documents (Zhu et al., 2021; Huang et al., 2020), post-editing model generated summaries for better factual consistency (Lee et al., 2022; Balachandran et al., 2022; Chen et al., 2021a), training summarization model with less noisy data by data filtering (Nan et al., 2021; Goyal and Durrett, 2021; Wan and Bansal, 2022a), and data augmentation-based methods (Wang et al., 2022b; Adams et al., 2022). The last category is usually combined with contrastive learning (Wan and Bansal, 2022b; Liu et al., 2021; Cao and Wang, 2021), which has shown a high effectiveness. However, contrastive learning often involves complex strategies to construct negative samples. For example, Cao and Wang (2021) use a combination of multiple methods including entity swapping, content masking and refilling, and low-confidence model generations.

Our work falls into the data augmentation and contrastive learning category. We adopt LLMs to construct negative samples with more diversity compared to previous strategies that have been predominantly driven by rules and heuristics.

2.2 Symbolic Knowledge Distillation

Symbolic knowledge distillation (West et al., 2022) is a conceptual framework originally proposed for constructing common-sense knowledge

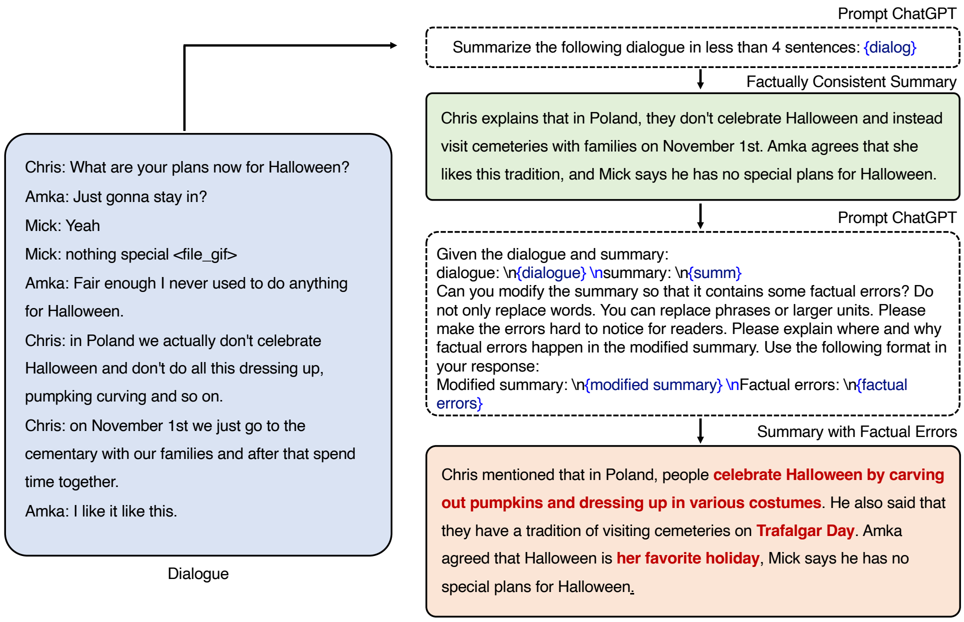

Figure 2: To extract symbolic knowledge from the teacher model (ChatGPT) for contrastive learning, we first prompt ChatGPT to generate a factually consistent summary, then use another prompt to instruct ChatGPT to modify the summary into a factually inconsistent version. The contents in red contain factual errors against the source dialogue.

graphs (Sap et al., 2019). A key advantage of the framework is that it does not require optimizing the student model on the teacher model's output probabilities, which was done in standard knowledge distillation (Hinton et al., 2015). Instead, it extracts symbolic knowledge (e.g., text) from the teacher model to construct a smaller student model.

Symbolic knowledge distillation has been used to construct better summarization models in different ways, motivated by the high-quality summaries generated by zero-shot and few-shot LLMs (Zhang et al., 2023), which are even preferred over human-written summaries. For example, Sclar et al. (2022) construct reference-free sentence summarization models with better controllability on the compression ratio, while Song et al. (2023) enhance summary abstractiveness via calibrated distillation. Liu et al. (2023c) use LLMs not only as a data augmenter to generate "quasi-references", but also as a summary evaluator to provide additional training signals. Jiang et al. (2024) distill LLM's summarization capability by generating multiple aspect-triple rationales and summaries, then utilize curriculum learning to train student models.

Our method differs from these studies by incorporating a stage that leverages both positive and negative summaries through contrastive learning to enhance the factual consistency of student models, while the studies above only consider positive examples.

3 Methodology

Given a dialogue D (aka "source documents" in document summarization studies), we aim to generate a summary S using a summarization model g that captures the main ideas of D. We specifically encourage S to be factually consistent with D, i.e., only including information directly found in D and not any information against the facts in D.

To construct more factually consistent and cost-effective dialogue summarization models, we first extract symbolic knowledge (i.e., augmented summaries) from a teacher model (ChatGPT), then use sequence-level knowledge distillation and contrastive learning to exploit the knowledge. An overview of our framework is shown in Figure 1.

3.1 Extracting Symbolic Knowledge

We use ChatGPT (*gpt-3.5-turbo*) to generate positive summaries which are supposed to be factually consistent with the source dialogue D, and negative summaries that contain factual errors against D. Specifically, we first prompt ChatGPT to generate k ($k = 3$) positive summaries for a dialogue, then we prompt it again to modify each positive summary into a negative one by modifying snippets of the summary (so we also have k negative summaries). An example is shown in Figure 2. We find that the quality of negative summaries improve when we explicitly prompt ChatGPT to explain the factual errors[2].

3.2 Utilising Symbolic Knowledge

The standard method to train summarization models is Maximum Likelihood Estimation (MLE). Specifically, given a single reference summary R^*, the summarization model g is encouraged to give the i-th token of R^* the maximum probability among all tokens in the vocabulary, based on the prefix string of the current token. The loss function, cross entropy, is defined as follows:

$$\begin{aligned} l_{mle} &= -\log(R^*|D) \\ &= -\sum_{i=1}^{n} \log P_g(R_i^*|D, R_{<i}^*) \end{aligned} \tag{1}$$

Here, R_i^* is the i-th token in R^*; $R_{<i}^*$ represents the tokens preceding R_i^*; and P_g is the probability distribution of the summarization model. Since there is only one reference summary, the loss function encourages the model to approximate the point mass distribution defined by the single reference (Liu et al., 2023c). As the loss function is defined at the word level in an autoregressive manner, it does not explicitly facilitate the factual consistency of the generated summary, which requires signals at semantic level and sequence level.

3.2.1 Sequence-level Distillation

Given that a large teacher model may generate more factually consistent summaries than the smaller student models, we employ Sequence-level Knowledge Distillation (SEQDISTILL) (Kim and Rush, 2016). This approach involves generating multiple quasi-summaries from the teacher model, which are then utilized as targets for fine-tuning the student models using cross-entropy loss. Given a set of positive summaries $\mathcal{P}^*$ generated by the teacher model, and the original human-written reference summary R^*, the loss function is as follows:

$$l_s = -\frac{1}{|\mathcal{P}^* \cup \{R^*\}|} \sum_{R \in \mathcal{P}^* \cup \{R^*\}} \log P_g(R|D)$$

The primary distinction between SEQDISTILL and Maximum Likelihood Estimation (MLE) lies in their method of distribution approximation. SEQDISTILL aims to approximate the teacher model's distribution, favoring multiple factually consistent summaries via a sampling-based method. Conversely, MLE approximates a point-mass distribution, where a single reference summary is given all the probability mass.

3.2.2 Contrastive Learning

We further incorporate two types of contrastive learning methods to boost the factual consistency of summarization models by incorporating negative summaries on top of SEQDISTILL.

Let $\mathcal{P}$ be a set of *positive summaries* that are factually consistent with the source dialogue D, $\mathcal{N}$ be a set of *negative summaries* that contain factual errors against D, and R be the target for cross entropy loss. A training instance with contrastive learning is a tuple $(D, R, \mathcal{P}, \mathcal{N})$. The loss function for a single training instance is defined as:

$$l = l_{mle} + \alpha \cdot l_c \tag{2}$$

where l_c is the contrastive loss, $\alpha \in [0, 1]$ is a hyperparameter to balance the two loss terms. Intuitively, l_c serves as a regularization term that shapes the distribution of the summarization model to favor factually consistent summaries. We employ two contrastive objectives, MARGINCONTRAST and PAIRCONTRAST, which differentiate between positive and negative summaries at the sequence and latent representation level, respectively.

MARGINCONTRAST aims to pull apart the positive summaries and negative summaries by enforcing a gap between sequence-level scores. Specifically, we aim to achieve higher scores for even the *worst positive summaries* than those of the *best negative summaries*, with the following loss:

$$l_c = \max\{0, \theta + \max\{S(\mathcal{N})\} - \min\{S(\mathcal{P})\}\} \tag{3}$$

Here, θ is the target score threshold, and $S(\cdot)$ is a scoring function. Inspired by BARTScore (Yuan et al., 2021), we define the scoring function $S(\cdot)$ for a summary X using the summarization model g as the length-normalized log-likelihood of all tokens:

$$S(X) = \frac{1}{m} \sum_{i=1}^{m} \log P_g(x_i|D, X_{<i}) \tag{4}$$

Here, m represents the number of tokens in X; x_i is the i-th token; and $X_{<i}$ are the preceding tokens.

[2] The average factual consistency (AlignScore) for 200 random positive summaries in the training set from the teacher model is 0.90 for SAMSum and 0.92 for DialogSum, indicating that positive summaries are mostly factually consistent. More details in Appendix A.2.

Dataset	#Train	#Dev	#Test	$\frac{\text{\#Speakers}}{\text{\#dial.}}$	$\frac{\text{\#Turns}}{\text{\#dial.}}$	$\frac{\text{\#Tokens}}{\text{dial.}}$
SAMSum	14,732	818	819	2.39	9.5	94
DialogSum	12,460	500	500	2.01	11.1	131

Table 1: Dataset statistics. **#Train**, **#Dev** and **#Test** refer to the numbers of dialogue-summary pairs (one summary per dialogue) in the training, development, and testing subsets. $\frac{\text{\#Speakers}}{\text{\#dial.}}$, $\frac{\text{\#Turns}}{\text{\#dial.}}$, and $\frac{\text{\#Tokens}}{\text{dial.}}$ refer to the average numbers of speakers, turns, and tokens in each dialogue.

Normalizing by m eliminates the impact of length on the evaluation of factual consistency.

PAIRCONTRAST differentiates positive from negative summaries by minimizing the similarities between their latent representations, while simultaneously maximizing the similarities among positive pairs. Let r_i, r_j, and r_k be summaries from either $\mathcal{P}$ or $\mathcal{N}$. We use $\mathbf{h_i}$ $\mathbf{h_j}$, and $\mathbf{h_k}$ to denote the vector-form representations of these summaries. The contrastive loss l_c is defined in accordance with the fomulation provided by Cao and Wang (2021) as follows:

$$l_c = -\frac{1}{\binom{|\mathcal{P}|}{2}} \sum_{\substack{r_i,r_j \in \mathcal{P} \\ r_i \neq r_j}} \log \frac{\exp(\mathrm{s}(\mathbf{h_i}, \mathbf{h_j})/\tau)}{\sum_{\substack{r_k \in \mathcal{P} \cup \mathcal{N} \\ r_k \neq r_i}} \exp(\mathrm{s}(\mathbf{h_i}, \mathbf{h_k})/\tau)} \tag{5}$$

Here, s is the *cosine* function; and τ is a temperature parameter (τ=1 in our experiments). We follow Cao and Wang (2021) to obtain the vector representations of the summaries by applying an MLP projection to the averaged last-layer outputs from the decoder for all tokens.

To summarize, MARGINCONTRAST uses summary log-likelihood estimated by the summarization model directly, while PAIRCONTRAST relies on the internal representation of summary words.

4 Experiment Setup

4.1 Datasets

We adopt two popular dialogue summarization datasets: SAMSum (Gliwa et al., 2019a) and DialogSum (Chen et al., 2021b). SAMSum is a collection of messenger-like conversations, while DialogSum contains daily conversations in a more real-life setting. In both datasets, there is one human-written reference summary for each conversation in the training split. Table 1 shows the statistics of the two datasets.

4.2 Student Models

We choose BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020) and Flan-T5 (Chung et al., 2024) as the student models, which have consistently demonstrated state-of-the-art performance in automatic text summarization (Zhao et al., 2022; Liu and Liu, 2021; Chung et al., 2024). Specifically, we use *facebook/bart-large*, *google/pegasus-large*, *google/flan-t5-large* as initial checkpoints. The number of learnable parameters for these models are 406 million, 568 million and 770 million, respectively, which are much smaller than that of the teacher model.

4.3 Baseline Models

FACTPEGASUS (Wan and Bansal, 2022a): an abstractive text summarization model for news summarization. It enhances factual consistency through several strategies: (1) factuality-oriented pre-training, (2) reference summary correction that addresses potential factual errors in reference summaries, (3) contrastive learning to boost the model's ability to differentiate between positive and negative summaries, where the negative summaries are constructed by rule-based entity swapping, (4) pre-training task simulation during fine-tuning that minimizes the gap between the pre-training and fine-tuning phases. We used their pre-trained model and code to fine-tune on our datasets.[3]

SWING (Huang et al., 2023): an abstractive dialogue summarization model that achieves state-of-the-art factual consistency and coverage on SAMSum and DialogSum. It leverages an uncovered loss to boost information coverage, and a contrastive loss to enhance factual consistency. We use their model generations directly.[4]

We also include the original human-written reference summaries (HUMANREF) to assess the relative quality compared to our method.

4.4 Evaluation Metrics

We selected multiple reference-free evaluation metrics, recognizing that our methods may produce high-quality summaries that diverge from human-written references. This divergence could lead to underrating by reference-based metrics. To assess factual consistency, we employed two state-of-the-art (SOTA) automatic metrics: an LLM-based metric, G-EVAL (Liu et al., 2023a), and a non-LLM-

[3] https://github.com/meetdavidwan/factpegasus
[4] https://github.com/amazon-science/AWS-SWING

	SAMSum							DialogSum						
	Const		UniEval			ROUGE		Const		UniEval			ROUGE	
Model	**S_A**	**S_G**	**Coh**	**Flu**	**Rel**	**R1**	**R2**	**S_A**	**S_G**	**Coh**	**Flu**	**Rel**	**R1**	**R2**
HUMANREF	0.80	4.80	0.92	0.93	0.97	1.00	1.00	0.82	4.84	0.94	0.92	0.98	1.00	1.00
Baselines														
FACTPEGASUS	0.63	3.08	0.87	0.90	0.73	0.45	0.20	0.67	3.44	0.88	0.87	0.77	0.49	0.24
SWING	0.82	4.38	0.93	0.93	0.84	0.52	0.28	0.83	4.54	0.95	0.93	0.90	0.53	0.29
MLE														
BART	0.82	4.27	0.92	0.93	0.84	**0.52**	**0.28**	0.80	4.22	0.94	0.93	0.88	**0.53**	**0.28**
PEGASUS	0.81	4.12	0.93	0.94	0.84	**0.50**	**0.26**	0.83	4.44	0.96	0.93	0.90	**0.52**	**0.28**
Flan-T5	0.82	4.34	0.93	0.93	0.84	**0.52**	**0.28**	0.84	4.65	**0.96**	**0.93**	**0.91**	**0.54**	**0.29**
SEQDISTILL (Our Method)														
BART	0.87	4.41	0.96	**0.94**	0.89	0.36	0.14	**0.93**	**4.81**	**0.98**	0.93	**0.93**	0.29	0.13
PEGASUS	**0.89**	**4.52**	0.95	**0.94**	**0.89**	0.39	0.17	0.90	**4.73**	**0.97**	0.93	**0.91**	0.42	0.22
Flan-T5	0.88	4.51	0.94	0.93	0.87	0.40	0.17	0.91	4.80	**0.96**	**0.93**	0.90	0.32	0.15
MARGINCONTRAST (Our Method)														
BART	0.89	**4.73**	0.97	**0.94**	0.90	0.40	0.18	**0.93**	4.72	**0.98**	**0.94**	**0.93**	0.31	0.15
PEGASUS	0.87	4.08	0.92	**0.94**	0.84	0.38	0.17	0.89	4.31	0.95	0.93	0.88	0.34	0.17
Flan-T5	0.90	4.69	0.95	**0.94**	0.88	0.42	0.20	0.91	4.76	0.95	**0.93**	0.90	0.37	0.19
PAIRCONTRAST (Our Method)														
BART	**0.91**	4.69	**0.98**	**0.94**	**0.92**	0.37	0.15	**0.93**	4.80	**0.98**	0.93	**0.93**	0.30	0.14
PEGASUS	**0.89**	4.47	**0.96**	**0.94**	**0.89**	0.38	0.16	**0.91**	4.62	0.96	**0.94**	**0.91**	0.36	0.18
Flan-T5	**0.91**	**4.74**	**0.96**	**0.94**	**0.90**	0.38	0.16	**0.93**	**4.86**	**0.96**	**0.93**	0.89	0.37	0.19

Table 2: Comparing different models and training strategies on Consistency (Const), Coherence (Coh), Fluency (Flu), Relevance (Rel) and ROUGE. We use two automatic factual consistency metrics, AlignScore (S_A) and G-Eval (S_G). Coherence, Fluency and Relevance are obtained from UniEval. R1 and R2 represent the F1 score of ROUGE 1 and ROUGE 2, respectively. We show the highest score(s) in all columns for the same model (e.g., BART) across {MLE, SEQDISTILL, MARGINCONTRAST, PAIRCONTRAST} in **bold** to show the most effective training strategy.

based metric, ALIGNSCORE (Zha et al., 2023b) [5]. This approach mitigates the potential bias of favoring LLM-generated summaries inherent in LLM-based metrics (Liu et al., 2023a). Additionally, we used UNIEVAL (Zhong et al., 2022a) to evaluate Coherence, Fluency, and Relevance. We also utilized the standard n-gram matching-based metric, ROUGE (Lin, 2004), primarily as a sanity check for models trained using MLE.

4.5 Other Experimental Details

For MARGINCONTRAST and PAIRCONTRAST, we merge the human-written reference R^* and positive summaries $\mathcal{P}^*$ generated by the teacher model as the positive set $\mathcal{P}' = \{R^*\} \cup \mathcal{P}^*$. For each training sample, we select one element $R \in \mathcal{P}'$ as the target for cross-entropy loss and use the rest as $\mathcal{P}$ for contrastive loss. All models are fine-tuned for 15,000 steps and evaluated at every 500 steps. The best checkpoint is selected according to AlignScore on the development set. We provide more implementation details in Appendix A.4.

[5] Our meta-evaluation on multiple dialogue summarization datasets show that AlignScore and G-Eval exhibit high correlation (0.4-0.7) with human evaluation results. More details in Appendix A.3.

5 Results and Discussions

5.1 The Effectiveness of Symbolic Knowledge Distillation and Contrastive Learning

We compare the performance of our methods (SEQDISTILL, MARGINCONTRAST and PAIR-

CONTRAST) and the baseline models on various quality dimensions, with a focus on factual consistency. From the results in Table 2, we make the following observations:

- Our distillation methods improve factual consistency (compared to baseline models and MLE methods) without sacrificing in other quality dimensions (i.e., Coherence, Fluency and Relevance).

- Our distillation methods consistently enhance the factual consistency of all pretrained models (BART, PEGASUS and Flan-T5). PAIRCONTRAST is generally the most effective method, although there is some performance variation depending on the dataset and pretrained model.

- SEQDISTILL and two contrastive learning methods result in significantly lower Rouge scores compared to MLE. However, it only tells us that there are fewer word overlaps between model generated summaries and human-written references rather than an actual quality decline. We will revisit this again with a case study in section 5.4.

- Flan-T5 in most cases generate more factually consistent summaries than BART and PEGASUS across different settings (MLE, SEQDISTILL, MARGINCONTRAST, PAIRCONTRAST).

- Flan-T5 with PAIRCONTRAST is the best summarization model overall, and it achieves comparable or sometimes better factual consistency, coherence and fluency than HUMANREF according to S_A, S_G and UNIEVAL.

5.2 The Effect of Human-written References

Observing that the best-performing student model demonstrates promising results, we further explore the impact of human-written references and seek to address the question: *Is it possible to construct dialogue summarization models without human-written references?*

Table 3 displays the performance of *flan-t5-large* trained using PAIRCONTRAST with various numbers of randomly sampled dialogues from the SAMSum training set. The quality scores on SAMSum test set across all dimensions are similar, whether original human-written reference summaries are employed (R=Y) or not (R=N), for all dataset sizes. These findings suggest the feasibility of developing robust summarization models using unlabeled datasets.

#Dialog	R^*	Const	Coh	Flu	Rel
300	N	0.89	0.96	0.93	0.88
300	Y	0.88	0.94	0.91	0.83
1000	N	0.89	0.94	0.92	0.86
1000	Y	0.89	0.95	0.93	0.86
3000	N	0.90	0.96	0.94	0.89
3000	Y	0.90	0.95	0.93	0.88
9000	N	0.91	0.96	0.93	0.88
9000	Y	0.90	0.96	0.94	0.89
13000	N	0.91	0.96	0.94	0.89
13000	Y	0.91	0.96	0.94	0.89

Table 3: Comparing the performance of *flan-t5-large* with PAIRCONTRAST on SAMSum, with ($R^* = Y$) or without ($R^* = N$) human-written references. $k = 3$ for all settings. The four quality dimensions are factual consistency (Const), coherence (Coh), fluency (Flu) and relevance (Rel). Factual consistency is obtained from AlignScore.

#Dialog	k	Consistency
1000	3	0.893
3000	1	0.898
3000	2	0.905
3000	3	0.902
9000	1	0.902
9000	2	0.904
9000	3	0.913

Table 4: Factual consistency (AlignScore) of *flan-t5-large* trained with PAIRCONTRAST on varying numbers of dialogues (#Dialog) and contrastive pairs per dialogue (k).

5.3 The Effect of the Number of Contrastive Pairs

Table 4 further shows the performance of *flan-t5-large* trained on different numbers of dialogues and contrastive pairs. We see that when the number of dialogues (i.e., #Dialog) is fixed, the model in general generates slightly more consistent summaries as k grows. On the other hand, there is no significant difference when we vary the number of contrastive pairs as long as the total number of training instances (i.e., #Dialog $\times$ k) is fixed. For example, when the total number of training instances is 9,000, (#Dialog=3000, k=3) yields the

Dialogue

Hannah: Hey, do you have Betty's number?
Amanda: Lemme check
Hannah: <file_gif>
Amanda: Sorry, can't find it.
Amanda: Ask Larry
Amanda: He called her last time we were at the park together
Hannah: I don't know him well
Hannah: <file_gif>
Amanda: Don't be shy, he's very nice
Hannah: If you say so..
Hannah: I'd rather you texted him
Amanda: Just text him 🙂
Hannah: Urgh.. Alright
Hannah: Bye
Amanda: Bye bye

FactPegasus (AlignScore=0.623)
Amanda doesn't have Betty's number but Hannah does. Larry called Betty last time they were at the park together.

MLE (AlignScore=0.766)
Hannah and Amanda are looking for Betty's number. Larry called Betty last time they were at the park. Amanda will text him.

SWING (AlignScore=0.888)
Hannah is looking for Betty's number. She doesn't know him well, but Amanda thinks she should ask Larry, who called Betty last time they were at the park together.

SeqDistill (AlignScore=0.902)
Hannah asks for Betty's number, but can't find it. She suggests asking Larry, who called her last time they went to the park together. However, she doesn't know Larry well and suggests that she should text him instead. They say goodbye.

Human-written Reference (AlignScore=0.907)
Hannah needs Betty's number but Amanda doesn't have it. She needs to contact Larry.

PairContrast (AlignScore=0.963)
Hannah asks Amanda for Betty's number, but Amanda can't find it and suggests asking Larry, who called Betty last time they were at the park together. Hannah is hesitant but Amanda encourages her not to be shy and to text Larry instead. Hannah agrees and says goodbye.

MarginContrast (AlignScore=0.980)
Hannah asks Amanda for Betty's number, but Amanda can't find it. Amanda suggests asking Larry, who called Betty last time they were at the park. Hannah is hesitant because she doesn't know Larry well but Amanda encourages her to do so. They end the conversation by saying goodbye.

Figure 3: An example dialogue from SAMSum (Gliwa et al., 2019a) with summaries generated by BART (Lewis et al., 2020) trained with different strategies (MLE, SEQDISTILL, MARGINCONTRAST, PAIRCONTRAST). Baseline models (FactPEGASUS, SWING) and human-written reference are included for comparison. Contents that are inconsistent with the input dialogue are shown in red. Ambiguous contents are shown in blue.

same result as (#Dialog=9000, k=1) does.

5.4 Case Study

Figure 3 presents an example dialogue along with summaries generated by different models, sorted by AlignScore (Zha et al., 2023b) in ascending order. The summaries from FACTPEGASUS, MLE, and SWING include factual errors unsupported by the dialogue. Specifically, FACTPEGASUS incorrectly asserts "but Hannah does" when in fact, Hannah does not have Betty's number. MLE inaccurately claims that "Hannah and Amanda are looking for Betty's number", though only Hannah is searching. In SWING's summary, "him" appears before the referent "Larry". For SEQDISTILL and Human-written reference, the pronouns "she" are ambiguous as there are multiple possible referent in previous context. Unlike these, summaries from PAIRCONTRAST and MARGINCONTRAST do not contain ambiguous references. Notably, our methods (SEQDISTILL, PAIRCONTRAST and MARGINCONTRAST) tend to produce longer summaries compared to the much more succinct human-written references, hence we see a substantially lower ROUGE scores for them (Table 2).

6 Conclusion

We investigated distilling LLM's symbolic knowledge (in the form of generated summaries) to enhance the factual consistency of smaller models for dialogue summarization. Our experiments with BART, PEGASUS, and Flan-T5 on the SAMSum and DialogSum datasets reveal that: (1) symbolic knowledge distillation enables the creation of more compact summarization models that surpass strong baselines which use complex data augmentation strategies; and (2) our best-performing student model, Flan-T5 with PAIRCONTRAST, produces summaries that are potentially better — in terms of factual consistency, coherence and fluency — than human-written references.

7 Limitations

The experiments in this paper are conducted on short daily dialogues. The findings may not generalize to other dialogue scenarios such as academic meetings and television interviews.

We use automatic evaluation metrics to assess the quality of model-generated summaries, which may not fully reflect human preferences.

8 Ethics Statement

This study is conducted under the guidance of the ACL code of Ethics.

References

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen Mckeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027.

Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. 2020. Knowledge distillation from internal representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7350–7357.

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830.

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.

Shuyang Cao and Lu Wang. 2021. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021a. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021b. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074.

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.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. Go figure: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019a. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *EMNLP-IJCNLP 2019*, page 70.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019b. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Kung-Hsiang Huang, Siffi Singh, Xiaofei Ma, Wei Xiao, Feng Nan, Nicholas Dingwall, William Yang Wang, and Kathleen Mckeown. 2023. Swing: Balancing coverage and faithfulness for dialogue summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 512–525.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Pengcheng Jiang, Cao Xiao, Zifeng Wang, Parminder Bhatia, Jimeng Sun, and Jiawei Han. 2024. Trisum: Learning summarization ability from large language models with structured rationale. *arXiv preprint arXiv:2403.10351*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

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.

Hwanhee Lee, Cheoneum Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Juae Kim, and Kyomin Jung. 2022. Factual error correction for abstractive summaries using entity retrieval. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 439–444.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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

Wei Liu, Huanqin Wu, Wenjing Mu, Zhen Li, Tao Chen, and Dan Nie. 2021. Co2sum: contrastive learning for factual-consistent abstractive summarization. *arXiv preprint arXiv:2112.01147*.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

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

Yixin Liu, Alexander R Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023c. On learning to summarize with large language models as references. *arXiv preprint arXiv:2305.14239*.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Mqag: Multiple-choice question answering and generation for assessing information consistency in summarization. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 39–53.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen Mckeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733.

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.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Melanie Sclar, Peter West, Sachin Kumar, Yulia Tsvetkov, and Yejin Choi. 2022. Referee: Reference-free sentence summarization with sharper controllability through symbolic knowledge distillation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9649–9668.

Hwanjun Song, Igor Shalyminov, Hang Su, Siffi Singh, Kaisheng Yao, and Saab Mansour. 2023. Enhancing abstractiveness of summarization models through calibrated distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7026–7036.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170.

David Wan and Mohit Bansal. 2022a. Factpegasus: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028.

David Wan and Mohit Bansal. 2022b. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022a. Analyzing and evaluating faithfulness in dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4897–4908.

Tianshu Wang, Faisal Ladhak, Esin Durmus, and He He. 2022b. Improving faithfulness by augmenting negative summaries from fake documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11913–11921.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023a. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023b. Alignscore: Evaluating factual consistency with a unified alignment function. In *Annual Meeting of the Association for Computational Linguistics*.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Conference of the European Chapter of the Association for Computational Linguistics*.

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.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2022. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Peng Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022a. Towards a unified multi-dimensional evaluator for text generation. In *Conference on Empirical Methods in Natural Language Processing*.

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

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733.

Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and detecting fine-grained factual errors for dialogue summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6825–6845.

A Appendix

A.1 Potential Risks

The summaries generated by ChatGPT may contain social biases, which require further investigation in real applications.

A.2 The Statistics and Quality of ChatGPT Summaries

We generated 3 positive and 3 negative summaries for 13,000 dialogues from the training split of SAMSum and 11,000 dialogues from the training split of DialogSum. For each dialogue, we made 6 API calls (3 for positive and 3 for negative) separately.

Table 5 shows the quality of 200 randomly sampled positive summaries generated by the teacher model *gpt-3.5-turbo*, validating that these summaries are mostly factually consistent, with high coherence, fluency and relevance as well.

Dataset	Const	Coh	Flu	Rel
SAMSum	0.90	0.97	0.94	0.91
DialogSum	0.92	0.97	0.94	0.94

Table 5: The factual consistency (Const), coherence (Coh), fluency (Flu) and relevance (Rel) for 200 randomly sampled positive summaries, generated by *gpt-3.5-turbo*, in the training set of SAMSum and DialogSum. Factual consistency is obtained from AlignScore (Zha et al., 2023b). Coherence, fluency, and relevance are obtained from UniEval (Zhong et al., 2022b).

A.3 Meta-evaluation of Factual Consistency Evaluation Metrics

We conducted a meta-evaluation of various automatic factual consistency metrics across three datasets: DiaSummFact (Zhu et al., 2023), FacEval (Wang et al., 2022a), and GO FIGURE (Gabriel et al., 2021). For the GO FIGURE dataset, we specifically utilized the subset derived from SAMSum (Gliwa et al., 2019a). In the case of DiaSummFact, we conducted evaluations at both the sentence level (DiaSummFact*) and summary level (DiaSummFact'). For the sentence-level evaluation, we excluded sentences whose labels include "Link Error" or "Coreference Error". All labels across the datasets were converted into a binary format: if any category of factual error is present, the label is marked as "factually inconsistent"; otherwise, it is marked as "factually consistent". The number of (dialogue, output) pairs in each dataset, where the output is either a sentence for sentence-level evaluation or a summary for summary-level evaluation, is presented in Table 6. Spearman and Pearson correlations are shown in Table 7 and Table 8.

Results show that both AlignScore and G-Eval exhibit high correlation with human annotations in most cases, except AlignScore on FacEval, which requires further investigation in future works. UniEval shows unsatisfactory correlation with human annotations on factual consistency, thus we only use AlignScore and G-Eval (*gpt-4*) for factual consistency evaluation.

A.4 Implementation Details

All models were fine-tuned for 15,000 steps with a batch size of 32 (per-device batch size 2/1, with gradient accumulation 16/32), evaluated every 500 steps (with model generations on development set) on an NVIDIA A100 GPU with 40G/80G memory. Each training task took between 4 to 72 hours,

	N
DiaSummFact*	475
DiaSummFact'	1240
FacEval	750
GO FIGURE	250

Table 6: The number of (dialogue, output) pairs (N) in the datasets for our meta-evaluation.

Metric	AlignScore	G-Eval	UniEval
DiaSummFact*	0.52	0.53	0.22
DiaSummFact'	0.48	0.60	0.15
FacEval	0.11	0.54	0.01
GoFigure	0.43	0.60	0.23

Table 7: Spearman correlation between automatic factual consistency evaluation metrics and human evaluation (binary).

depending on the size of the model.

We searched for the best hyper-parameters of $\alpha \in \{0.5, 1, 2\}$ for PAIRCONTRAST, and $\alpha \in \{0.5, 1, 2\}$ and $\theta \in \{15, 30\}$ for MARGINCONTRAST, according to AlignScore (Zha et al., 2023b) on development set.

The code for PAIRCONTRAST was developed based on CLIFF [6]. ROUGE scores are computed using Python package **evaluate 0.4.0** with default parameters [7].

A.5 License or Terms

Our code and data will be released under MIT license.

A.6 Intended Use of Existing Artifacts

The SAMSum dataset, as presented in Gliwa et al. (2019b), is distributed under the Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license. We offer supplementary details (e.g., model-generated summaries), while preserving the integrity of the original data, comprising dialogues and reference summaries.

A.7 Artifacts

The artifacts we release (code, data) are all in English only.

Metric	AlignScore	G-Eval	UniEval
DiaSummFact*	0.49	0.54	0.17
DiaSummFact'	0.39	0.49	0.13
FacEval	0.09	0.49	-0.01
GoFigure	0.44	0.71	0.23

Table 8: Pearson correlation between automatic factual consistency evaluation metrics and human evaluation (binary).

[6] https://github.com/ShuyangCao/cliff_summ/tree/main/models

[7] https://pypi.org/project/evaluate/