Semi-Supervised Dialogue Abstractive Summarization via High-Quality Pseudolabel Selection

Jianfeng He[1,2*], **Hang Su**[1], **Jason Cai**[1], **Igor Shalyminov**[1], **Hwanjun Song**[1], **Saab Mansour**[1]
[1] AWS AI Labs
[2] Virginia Tech
jianfenghe@vt.edu, {shawnsu, cjinglun, shalymin, saabm}@amazon.com

Abstract

Semi-supervised dialogue summarization (SSDS) leverages model-generated summaries to reduce reliance on human-labeled data and improve the performance of summarization models. While addressing label noise, previous works on semi-supervised learning primarily focus on natural language understanding tasks, assuming each sample has a unique label. However, these methods are not directly applicable to SSDS, as it is a generative task, and each dialogue can be summarized in different ways. In this work, we propose a novel scoring approach, SiCF, which encapsulates three primary dimensions of summarization model quality: Semantic invariance (indicative of model confidence), Coverage (factual recall), and Faithfulness (factual precision). Using the SiCF score, we select unlabeled dialogues with high-quality generated summaries to train summarization models. Comprehensive experiments on three public datasets demonstrate the effectiveness of SiCF scores in uncertainty estimation and semi-supervised learning for dialogue summarization tasks. Our code is available at `https://github.com/amazon-science/summarization-sicf-score`.

1 Introduction

Dialogue summarization generates concise summaries of dialogues, helping users quickly understand key points without navigating through complex contexts (Feng et al., 2021). This study prioritizes abstractive summarization, which offers more flexibility than extractive approaches (Gupta and Gupta, 2019; Wong et al., 2008). Despite its wide applicability in scenarios like meetings and casual conversations, dialogue summarization faces challenges such as scarcity of annotations and high annotation costs. However, the proliferation of pretrained models and unlabeled dialogues offers a solution. In this paper, we explore **S**emi-**S**upervised **D**ialogue **S**ummarization (SSDS) (Chen and Yang, 2021), aiming to enhance dialogue summarization by a small labeled dataset alongside a large collection of unlabeled dialogues.

Previous SSDS research (Chen and Yang, 2021) has used data augmentation to increase the size of both labeled and unlabeled dialogue datasets, but the issue of pseudolabel noise has been largely overlooked. Specifically, an initial model fine-tuned on labeled samples is used to generate pseudolabels for unlabeled samples. Then, the unlabeled samples and their pseudolabels are used to train the semi-supervised model (Rizve et al., 2021). However, the imperfections of the initial model can lead to pseudolabel noise, such as hallucination and missing key information. Pseudolabel noise is a big concern in semi-supervised learning because training on pseudolabels with significant noise can deteriorate model performance. Thus, we aim to address pseudolabel noise in SSDS in this research.

Many existing solutions for pseudolabel noise estimation and mitigation are designed for understanding tasks (e.g., classification (Cordeiro and Carneiro, 2020)), and they are not directly applicable to SSDS due to inherent diversity of ground truth summaries. Specifically, these solutions, such as Mix-Up in Mix-Match (Berthelot et al., 2019), assume each sample has a unique label, representing a single attribute like a semantic class. In contrast, SSDS is a generation task where each dialogue can be summarized in different ways. For instance, summaries like "the audience is happy to hear the news" and "the news makes the audience glad" convey the same message with different wording. As a result, SSDS, like other generation tasks, does not have a unique label per sample. This distinction makes previous pseudolabel noise solutions unsuitable for SSDS. Thus, we need a new and generalized pseudolabel noise measurement solution that considers label diversity in SSDS, with-

[*] The work was done during an AWS AI Labs internship.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5976–5996
June 16-21, 2024 ©2024 Association for Computational Linguistics

out relying on ground truth summaries, as unlabeled dialogues lack them.

To achieve this, we propose assessing pseudolabel quality: a predicted summary with higher quality indicates less noise in pseudolabels. We thus propose the SiCF score, which measures summary quality based on common characteristics of high-quality summaries, such as model confidence, information coverage, and faithfulness to the original dialogues. SiCF comprises three components: "Semantic invariance" assesses model confidence at the text level, "Coverage" evaluates key information captured at the word level, and "Faithfulness" measures alignment with the original dialogues at the sentence level. As shown in Figure 1, we then rank and select the unlabeled dialogues with high-quality pseudolabels as indicated by the SiCF score. Besides, since uncertainty estimation is a representative way to estimate the model prediction quality (Gawlikowski et al., 2023; He et al., 2023b) and Bayesian Neural Network (BNN) is an effective uncertainty estimation method (Mukhoti et al., 2023), we propose a variant-length multi-label BNN for our SiCF score. Our contributions are as below.

- We propose the SiCF score framework to measure the quality of predicted summaries based on these three key characteristics. To the best of our knowledge, we are the first to comprehensively evaluate summary quality without relying on ground truth summaries.
- We introduce a variant-length multi-label BNN uncertainty estimation technique used in the SiCF score. In contrast, conventional BNN (Mukhoti et al., 2023) is designed for fixed-length single-label cases, which do not align with the requirements of our task.

2 Related Work

Semi-supervised text summarization. Please refer to Sec. A.1 for the related work of Semi-supervised text summarization.

Semi-supervised dialogue summarization. Semi-supervised dialogue summarization is also underexplored, although some works focus on guiding dialogue summarization (Liu and Chen, 2021), improving model performance via human feedback (Chen et al., 2022), and enhancing factual consistency between ground-truth and generated summaries (Chen et al., 2021a).

In terms of semi-supervised extractive dialogue summarization, Mishra et al. (2023) employ GPT 3.5 for quality assessment based on token probabilities. Zhuang et al. (2023) introduce self-supervised pre-training to enhance BERT's ability to contextualize dialogue representations.

Regarding our focus, that is the semi-supervised abstractive dialogue summarization, CODA (Chen and Yang, 2021) is proposed to address SSDS using data augmentation. While data augmentation can expand the size of both labeled and unlabeled data, it overlooks challenges posed by pseudolabel noise, a prevalent issue in semi-supervised learning. Unlike previous SSDS models that overlooked pseudolabel noise, our goal is to enhance SSDS performance by measuring pseudolabel quality and effectively eliminating unreliable pseudolabels.

Solution of pseudolabel noise in semi-supervised learning. Many methods have been proposed for label noise in natural language understanding tasks (Cordeiro and Carneiro, 2020; Berthelot et al., 2019; He et al., 2023a; Lei et al., 2022). However, most of these methods are not directly applicable to SSDS, because this generation task has diverse ground truth summaries for each dialogue. While some of these methods have the potential to be applied towards SSDS, like teacher-student knowledge distillation model for noisy text summarization (Liu et al., 2020), they do not consider the diversity of summaries within SSDS. Rizve et al. (2021) have a similar task setting to ours, but their task is multi-label image classification, which still provides a unique label for each image. Wan et al. (2023) focus solely on model generation without considering the interaction with context (e.g., dialogue in our task). In contrast, we consider both model prediction itself by semantic invariance and the relation between generations and context via coverage and faithfulness.

As for injecting noise into the dialogues or pseudolabels (He et al., 2019), it focuses on improving the model's robustness through training with this injected noise and aims to mitigate the impact of noise. In contrast, our work aims to measure the extent of sample noise. Furthermore, they need to retrain the model, and the added noise might degrade the model's performance. In contrast, our work does not require retraining the model and will not harm its performance.

Summary quality. Please refer to Sec. A.1 for the related work of "Summary quality".

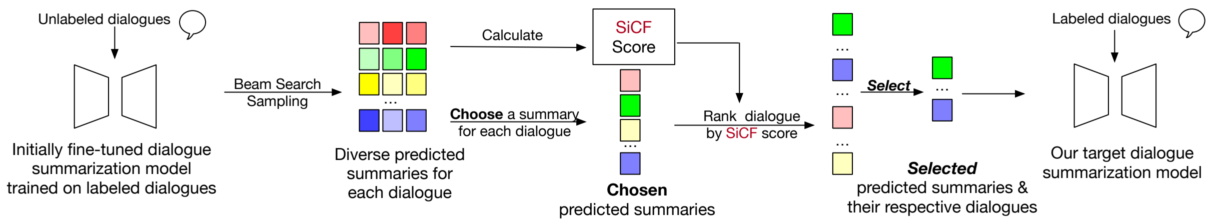

Unlabeled dialogues
Initially fine-tuned dialogue summarization model trained on labeled dialogues
Beam Search Sampling
Diverse predicted summaries for each dialogue
Calculate
SiCF Score
Choose a summary for each dialogue
Chosen predicted summaries
Rank dialogue by SiCF score
Select
Selected predicted summaries & their respective dialogues
Labeled dialogues
Our target dialogue summarization model

Figure 1: A global view of our SSDS framework using the semantic invariance, coverage, and faithfulness combined score (SiCF). Each row in the colored matrix represents diverse predicted summaries for a dialogue. For each unlabeled dialogue, the predicted summary closest to mean embedding is chosen. We then rank the chosen predicted summaries by the SiCF scores and select a portion of them. The selected <unlabeled dialogues, pseudolabels> and all human-labeled pairs are used for our target model learning. The detailed SSDS framework is outlined in Sec. 4.1.

3 Problem Setting

We are given a dialogue set D^l with annotated summaries and a dialogue set D^u without annotations. D^l and D^u belong to the same domain, as our focus is not domain generalization. Given a pretrained dialogue summarization model G_0, we leverage D^l and some or all data of D^u to fine-tune G_0, obtaining a target dialogue summarization model $\hat{G}$. We aim to accurately evaluate the quality of pseudolabels for unlabeled samples, enabling us to select higher-quality unlabeled data for training an improved model $\hat{G}$.

4 Our Model

4.1 Overview of SSDS via our SiCF Score

The proposed framework for SSDS with SiCF scores is shown in Figure 1. We begin with a pretrained dialogue summarization model G_0, such as DialogLED (Zhong et al., 2022a) [1]. We first fine-tune G_0 on dialogue-summary pairs in D^l. To assess pseudolabel quality, we use uncertainty estimation, which is an effective way to measure the model prediction quality (Gawlikowski et al., 2023). Bayesian Neural Network (BNN) is an effective method for uncertainty estimation and is often approximated by ensemble (Gal and Ghahramani, 2016). We generate k diverse summaries for each unlabeled dialogue from D^u by beam search sampling (Vijayakumar et al., 2016). Next, based on these diverse summaries, we calculate our SiCF score for each unlabeled dialogue to evaluate its summary quality. This score includes three aspects: Semantic invariance, Coverage, and Faithfulness. Moreover, we **choose** a summary for each dialogue based on the embedding that is closest to the mean of its all diverse summary embeddings. Next, we rank and **select** high-quality dialogue-pseudolabel pairs based on the SiCF scores. Finally, we fine-tune G_0 with the labeled dialogues and the *selected* unlabeled dialogues with pseudolabels to train the target dialogue summarization model $\hat{G}$.

Our work focuses on (1) how to obtain SiCF scores that accurately measure the quality of predicted summaries based on uncertainty estimation, and (2) how to use SiCF scores to select high-quality unlabeled dialogues and then improve $\hat{G}$.

We detail the reasons for choosing semantic invariance, coverage, and faithfulness in Sec. A.2.1.

4.2 SiCF Score: Semantic Invariance

Kuhn et al. (2023) propose semantic uncertainty based on semantic invariance for text generation quality evaluation. Higher semantic invariance means smaller semantic divergence between the k diverse generations of a sample, and thus indicating a higher quality in generations' semantics. However, Kuhn et al. (2023) needs to cluster diverse generations for each sample, which is time-consuming for a large sample size.

Different from Kuhn et al. (2023), we propose a variance-based method to measure the semantic invariance without clustering. This is because variance is also an effective uncertainty estimation method when the task has no unique ground truth (e.g., text summarization) (Chen, 2019). Specifically, given a dialogue from D^u with k diverse predicted summarizations $s = \{s_1, s_2, ..., s_k\}$, we use a pretrained encoder model (e.g., RoBERTa (Liu et al., 2019)) to produce their text embeddings as $e_1, e_2, ..., e_k$, we get a semantic invariance score λ_{SeIn} for the unlabeled dialogue by the variance of its diverse summary embeddings as follows,

$$\lambda_{SeIn} = var(e_1^s, e_2^s, ..., e_k^s) \qquad (1)$$

[1] We select DialogLED for its superior performance and hardware efficiency, facilitating accessibility for other researchers using our setup.

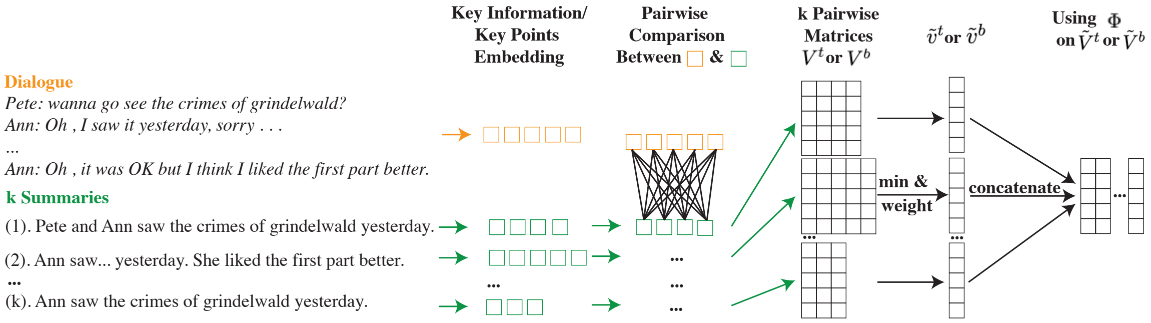

Key Information/
Key Points
Embedding
Pairwise
Comparison
Between □ & □
k Pairwise
Matrices
V^t or V^b
$\tilde{v}^t$ or $\tilde{v}^b$
Using Φ
on $\tilde{V}^t$ or $\tilde{V}^b$
Dialogue
Pete: wanna go see the crimes of grindelwald?
Ann: Oh , I saw it yesterday, sorry . . .
...
Ann: Oh , it was OK but I think I liked the first part better.
k Summaries
(1). Pete and Ann saw the crimes of grindelwald yesterday.
(2). Ann saw... yesterday. She liked the first part better.
...
(k). Ann saw the crimes of grindelwald yesterday.
min &
weight
concatenate

Figure 2: The global view of our coverage and faithfulness scores in our SiCF score.

Our variance operation is more efficient than Kuhn et al. (2023), as variance achieves a time complexity of $O(k)$ compared to their $O(k^2)$.

4.3 SiCF Score: Coverage

Since a good summary typically covers key details in a dialogue, we design a coverage score to measure the quality of a summary. Unlike coverage in Huang et al. (2023), ours does not depend on ground truth summaries.

To get our coverage score, we need to extract key details from a dialogue. Based on our observation (illustrated in Sec. A.2.2) and conciseness of summaries, we use the nouns in a dialogue to represent its key details. This is because the nouns in a dialogue carry the information to distinguish themselves from other dialogues.

Concretely, we use a pretrained POS tagging model (i.e., Flair (Akbik et al., 2019)) to extract nouns from a dialogue as its key details. Also, key details (nouns) can be extracted from a corresponding summary for comparison. We use the $T^d = [t^d_1, t^d_2, ..., t^d_p]$ to represent a sequence of noun embedding from a dialogue d, which has p nouns. Similarly, we use the $T^s = [t^s_1, t^s_2, ..., t^s_q]$ to represent a sequence of noun embedding of a summary s. As shown in Figure 2, we then calculate the similarity matrix $V^t \in \mathbb{R}^{p\times q}$ between noun embeddings T^d and T^s,

$$V^t = Dist(T^d, T^s) \tag{2}$$

where $Dist$ is pair-wise Euclidean distance. We choose Euclidean distance, as we expect a small value to mean high quality. A smaller value in V^t means better similarity between a noun of the dialogue and a noun of its one summary. Thus, we then apply row-level min operation on V^t to get coverage vector $\hat{v}^t \in \mathbb{R}^p$, that is $\hat{v}^t = min(V^t)$, where each element indicates the coverage degree between a predicted summary and a noun in the dialogue. We further weight the $\hat{v}^t$ as $\tilde{v}^t = w^t \cdot \hat{v}^t$, where $\tilde{v}^t \in \mathbb{R}^p$. The $w^t \in \mathbb{R}^p$ is the weight of each noun in dialogue, measured by noun's occurrence. Since speaker names in dialogues are proper nouns that often repeat, we take the maximum occurrence of proper nouns as 1 to prevent bias in the model due to speaker names.

Since we have k diverse generated summaries for each dialogue, we can have $\tilde{V}^t = [\tilde{v}^t_1, \tilde{v}^t_2, ..., \tilde{v}^t_k] \in \mathbb{R}^{k\times p}$, where $\tilde{v}^t_i$ is a coverage vector of i-th diverse generated summary for the dialogue. Since the coverage score should be a scalar, we use a function Φ to get a coverage score λ_{cov},

$$\lambda_{cov} = \Phi(\tilde{V}^t) \tag{3}$$

where Φ can be mean, BNN, or their combination (m+BNN), which will be introduced in Sec. 4.5.

4.4 SiCF Score: Faithfulness

Because a good summary should adhere to the key point of the dialogue, we consider faithfulness, which is the adherence degree between the key points of a dialogue and its summaries. However, using details (e.g., nouns) as key points may omit the connection of state words, like "not" and "disagrees". But using the text-level embedding is too general to miss the fine-grained information, such as SummaC (Laban et al., 2022). As a result, we consider sentence-level key points, because it keeps both state words and fine-grained information. Unlike faithfulness in Huang et al. (2023), ours does not depend on ground truth summaries.

Specifically, given a dialogue with h sentences and a predicted summary with z sentences, we have a sequence of dialogue sentence embeddings $B^d = [b^d_1, b^d_2, ..., b^d_h]$ and a sequence of summary sentence embeddings $B^s = [b^s_1, b^s_2, ..., b^s_z]$ for them by an encoder of a pretrained Natural Language Inference

(NLI) model, which is effective in faithfulness-check models (e.g., FactCC (Kryscinski and McCann, 2021) and SummaC (Laban et al., 2022)). As shown in Figure 2, we utilize the pretrained NLI model to obtain a faithfulness matrix $V^b \in \mathbb{R}^{h \times z}$ as follows.

$$V^b = NLI(B^d, B^s) \tag{4}$$

The $h \times z$ shape is built by pair-wise comparing h sentences in a dialogue to z sentences in a summary. Each element in $NLI(B^d, B^s)$ is obtained by first calculating the NLI negative and positive results between i-th dialogue sentence embedding b_i^d and j-th summary sentence embedding b_j^s, followed by returning the NLI result of these two sentences. The NLI result in our work is the negative score subtracting the positive score, which is similar to SummaC. As a result, a smaller element in V^b means better faithfulness between a dialogue sentence and a summary sentence.

Next, similar to the coverage score, we apply row-level min operation on V^b and have $\hat{v}^b \in \mathbb{R}^h$. Each element in $\hat{V}$ indicates the faithfulness agreement between the summary sentences and a dialogue sentence. We further weight $\hat{v}^b$ as $\tilde{v}^b = \hat{v}^b \cdot w^b$, where $w^b \in \mathbb{R}^h$ has each element as the noun occurrences in a dialogue sentence. We also limit the proper noun occurrences to a maximum of 1 because names in the dialogue are frequently mentioned and less significant than other nouns.

Since there are k summaries for each dialogue, we can then obtain $\tilde{V}^b = [\tilde{v}_1^b, \tilde{v}_2^b, ..., \tilde{v}_k^b] \in \mathbb{R}^{k \times h}$. Finally, similar to Eq. 3, the faithfulness score for the unlabeled dialogue is as follows,

$$\lambda_{fai} = \Phi(\tilde{V}^b) \tag{5}$$

Operation Φ will be introduced in Sec. 4.5.

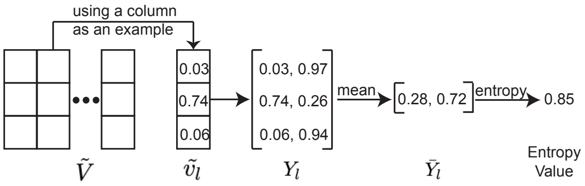


Figure 3: The diagram of the variant-length multi-label BNN. It uses a $\tilde{V}$ column as an example to obtain an entropy value. This example sets $k = 3$. The $\lambda_{cov/fai}$ is the sum of the entropy values from all $\tilde{V}$ columns.

4.5 Mean and Bayesian Neural Network

Once we have a coverage matrix $\tilde{V}^t \in \mathbb{R}^{k \times p}$ or a faithfulness matrix $\tilde{V}^b \in \mathbb{R}^{k \times h}$, we propose three types of operations Φ to calculate coverage score or faithfulness score, which all measure prediction quality. For simplicity, we let $\tilde{V}$ denote $\tilde{V}^t$ or $\tilde{V}^b$.

Mean. As a straightforward method (Zhang et al., 2024), we consider the mean value of $\tilde{V}$ to be the required scalar score, that is, coverage or faithfulness score $\lambda_{cov/fai} = mean(\tilde{V})$. In this case, $\lambda_{cov/fai}$ represents the average score across all k diverse summaries, measuring coverage or faithfulness.

Multi-label BNN. However, the mean operation only considers the values themselves but ignores the distribution between $[\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_k]$. BNN has been an effective distribution-based uncertainty estimation method and is suitable for quality assessment (Gawlikowski et al., 2023). However, BNN is originally designed for the fixed-length single-label case, while our case is for variant-length multi-label. Concretely, each element in $\tilde{v}$ can be taken as a label (a noun or a key point) and a good summary should belong to each label, thus our case is multi-label. Besides, each dialogue might have different sizes of nouns (p) and key points (h), and hence our case has variant lengths. Thus, we propose a multi-label BNN to consider the distribution as below,

$$\lambda_{cov/fai} = \sum_{l=1}^{p\ (or\ h)} \mathbb{H}[Y_l | x, D] \tag{6}$$

where we treat the multi-label problem as p or h binary single-label problem. Concretely, we first conduct min-max normalization on $\tilde{V}$. Then, shown as in Figure 3, we get $Y_l \in \mathbb{R}^{k \times 2}$ by $\tilde{v}_l \in \mathbb{R}^k$ which is a column of the $\tilde{V}$. Y_l is a concatenation of k vector $[\tilde{V}_{i,l}, 1 - \tilde{V}_{i,l}]$ for i-th predicted summary in l-th label. Eq. 6 first calculates the mean $\bar{Y}_l \in \mathbb{R}^{1 \times 2}$ of $Y_l \in \mathbb{R}^{k \times 2}$ for l-th label, followed by calculating the entropy of $\bar{Y}_l$. Finally, the sum of all p or h labels' entropy values is $\lambda_{cov/fai}$. We use sum rather than mean because our task is variant length and we expect all labels' uncertainty to be accumulated. More detail is in Sec. A.2.3.

m+BNN. The mean operation only considers the prediction quality from the view of logits. The BNN method only considers the prediction quality from the view of uncertainty estimation. Therefore, we propose m+BNN to multiply them as below,

$$\lambda_{cov/fai} = mean(\bar{V}) \times \sum_{l=1}^{p\ (or\ h)} \mathbb{H}[Y^l | x, D] \tag{7}$$

Table 1: Table of data splitting. The numbers in brackets are the sample size.

	Data split setting with small-size labeled data		Data split setting with medium-size labeled data	
	Labeled data size	Unlabeled data size	Labeled data size	Unlabeled data size
SAMSUM	1% (147)	50% (7366)	5% (736)	50% (7366)
DIALOGSUM	1% (124)	50% (6230)	5% (623)	50% (6230)
TODSUM	2% (157)	90% (7103)	10% (789)	90% (7103)

4.6 Fine-Tune on Unlabeled Dialogues Selected by SiCF Scores

Once we have obtained λ_{sein}, λ_{cov}, and λ_{fai}, we would like to merge them into a SiCF score. However, these three scores are on different scales. Therefore, we use the permutation of pseudolabels based on each of these three scores.

Concretely, given a $\lambda \in \{\lambda_{sein}, \lambda_{cov}, \lambda_{fai}\}$, we sort the pseudolabels in descending order based on λ. As a result, an order number δ is positively correlated to the quality of pseudolabel, because a lower λ means higher quality. To calculate the SiCF score, we then proceed as follows,

$$\lambda_{SiCF} = (\alpha\delta_{SeIn} + \beta\delta_{cov} + \gamma\delta_{fai})/3N \quad (8)$$

where α, β, and γ are hyperparameter. A larger λ_{SiCF} means higher quality. As shown in Figure 1, we select a certain ratio (e.g., 25%) of unlabeled dialogue by the SiCF scores on the chosen predicted summaries, which is described in Sec. 4.1. Finally, we fine-tune G_0 on selected <unlabeled dialogues, pseudolabel> and all labeled pairs to get our target dialogue model $\hat{G}$.

5 Experiments

5.1 Task setting

Our goal is to select a certain ratio of <unlabeled dialogues, pseudolabels> with high SiCF scores, which are used to fine-tune the baseline model together with all labeled samples. In addition, we are interested in the effectiveness of SiCF score from a view of uncertainty estimation.

5.2 Data

We conduct experiments on three datasets: (1) SAMSUM (Gliwa et al., 2019) is a daily chat domain dataset with 14732 training samples, 818 evaluation samples, and 819 testing samples. (2) DIALOGSUM (Chen et al., 2021b) is a dataset in real-life scenarios with training, evaluation, and testing sample sizes of 12460, 500, and 500, respectively. (3) TODSUM (Zhao et al., 2021) is a dataset with task-oriented dialogues in seven domains with training, evaluation, and testing sample sizes of 7460, 999, and 999, respectively. We split each dataset into labeled and unlabeled portions, and experiment with two settings (**small** and **medium**-size labeled setting) controlling for the size of the labeled data as shown in Table 1.

5.3 Baselines

We compare our proposed method to two baselines: (1) **Random Rank**: a method using the same chosen sampling ratio but using a random rank, (2) **Full Unlabeled**: a method using all unlabeled dialogues without any selection, which has been verified to be a strong baseline in selective learning (He et al., 2023a). We also calculate an upper bound for ranking termed **pseudo oracle**, where the ground truth summary is used to score the pseudo summaries according to the BERTScore-F.

5.4 Metrics

Metric for uncertainty estimation, force-truth evaluation, evaluates the quality score from a view of uncertainty estimation (Zhang et al., 2019b; He et al., 2020, 2023c). Concretely, it simulates the performance improvement of quality scores with human involvement. We measure F-values of BERTScore, ROUGE-1, 2, and L at different elimination ratios. Concretely, for N testing samples and an elimination ratio r, the most uncertain predicted summaries in size of $N \times r$ are set as ground truth summaries. The more accurate the uncertainty scores we obtain, the more inaccurate predicted summaries will be replaced by ground truth summaries under the same r, resulting in a better summarization metric (e.g., ROUGE-1) score. The dialogue summary metric score at a 0% eliminated ratio represents the original model's summarization performance. The summary metric scores at 10%-90% elimination ratios reflect the uncertainty estimation results. We report the mean of metric performance between 0-50% and between 0-90%. **Metrics for SSDS** are BERTScore (Zhang et al., 2019a), ROUGE-1, -2, and -L (Lin, 2004). Besides, since we have calculated pseudo oracle, we can get **SSDS improved ratios**, which indicates the improved degree compared to pseudo oracle. A SSDS

Table 2: Uncertainty estimation results of SiCF score on three datasets in terms of BERTScore-F. The "0-50" and "0-90" are the mean of BERTScore-F with eliminated ratio range 0%-50% and 0%-90%. In the medium-size setting, the TODSUM has a mean and standard deviation as shown in Table 14.

ROUGE-1	SAMSUM(1:50)		DIALOGSUM(1:50)		TODSUM(2:90)		SAMSUM(5:50)		DIALOGSUM(5:50)		TODSUM(10:90)	
	0-50%	0-90%	0-50%	0-90%	0-50%	0-90%	0-50%	0-90%	0-50%	0-90%	0-50%	0-90%
Random Rank	57.92	69.13	58.03	69.21	77.66	83.65	59.39	70.21	59.46	70.24	81.05	86.11
SiCF(mean)	58.90	70.28	58.85	70.08	78.78	84.80	60.45	71.44	60.42	71.29	82.05	87.17
SiCF(BNN)	59.34	70.64	**59.57**	**70.82**	78.51	84.64	60.87	71.80	**61.03**	**71.93**	82.08	87.27
SiCF(mean+BNN)	59.38	70.69	59.25	70.46	78.85	84.91	60.80	71.74	60.78	71.68	82.23	87.38
SiCF(m+BNN-s)	**59.47**	**70.78**	59.37	70.61	**78.95**	**84.97**	**60.95**	**71.89**	60.74	71.68	**82.27**	**87.43**
Pseudo Oracle	61.35	72.88	61.59	72.98	81.56	87.70	62.95	74.07	62.98	74.00	84.98	90.16

Table 3: SSDS results on SAMSUM and TODSUM. The values in the bracket are SSDS improved ratios. ROUGE-L is listed in Tables 6 and 7. In the medium-size setting, the TODSUM has a mean and standard deviation as shown in Table 15. The SSDS results on DIALOGSUM are listed in Table 8.

	Small-Size Labeled Data			Medium-Size Labeled Data		
	ROUGE-1	ROUGE-2	BERTScore-F	ROUGE-1	ROUGE-2	BERTScore-F
			SAMSUM			
Initial Fine-Tuned	43.90(0%)	18.49(0%)	43.74(0%)	46.81(0%)	20.46(0%)	45.74(0%)
Full Unlabeled	44.32(41%)	19.07(42%)	43.72(-3%)	47.76(62%)	21.84(91%)	**47.02(81%)**
Random Rank	44.98(105%)	19.32(60%)	44.39(112%)	47.65(55%)	21.13(44%)	46.66(58%)
SiCF (mean)	**45.85(191%)**	19.90(102%)	**44.89(198%)**	47.90(71%)	21.39(61%)	46.51(48%)
SiCF (BNN)	45.20(127%)	**19.95(105%)**	44.45(122%)	47.77(63%)	**22.07(106%)**	46.35(38%)
SiCF (m+BNN)	45.14(121%)	19.31(59%)	44.47(125%)	**48.14(87%)**	22.05(105%)	46.61(55%)
SiCF (m+BNN-s)	45.40(147%)	19.38(64%)	44.34(103%)	47.83(67%)	21.40(62%)	46.55(51%)
Pseudo Oracle	44.92(100%)	19.87(100%)	44.32(100%)	48.33(100%)	21.97(100%)	47.32(100%)
			TODSUM			
Initial Fine-Tuned	76.60(0%)	59.51(0%)	70.02(0%)	79.75(0%)	64.69(0%)	74.28(0%)
Full Unlabeled	77.02(16%)	59.86(9%)	70.57(13%)	80.00(6%)	64.99(5%)	74.30(0%)
Random Rank	76.91(12%)	59.43(-2%)	70.64(15%)	80.57(22%)	65.73(19%)	75.13(20%)
SiCF (mean)	**77.94(53%)**	**61.01(40%)**	**71.72(43%)**	81.08(36%)	66.70(37%)	75.75(34%)
SiCF (BNN)	76.64(1%)	59.66(4%)	70.53(12%)	**82.01(61%)**	**67.90(59%)**	**76.74(58%)**
SiCF (m+BNN)	77.01(16%)	59.92(11%)	70.91(22%)	80.93(32%)	66.32(30%)	75.78(35%)
SiCF (m+BNN-s)	76.45(-6%)	59.46(-1%)	70.64(15%)	81.22(39%)	67.14(45%)	76.15(44%)
Pseudo Oracle	79.09(100%)	63.19(100%)	73.96(100%)	83.43(100%)	70.10(100%)	78.48(100%)

improved ratio (IR) is defined as below, which is similar to normalized WER in Gu et al. (2023),

$$IR = \frac{MS_m - MS_{ini}}{MS_{ora} - MS_{ini}} \tag{9}$$

where MS_m is a method's metric score, MS_{ini} is the initial finie-tuned's metric score, and MS_{ora} is the pseudo oracle's metric score. A higher improved ratio signifies method superiority, further explained in Sec. A.3.1.

5.5 Experimental Setting

We first only use all labeled dialogues to fine-tune the pretrained DialougeLED (Zhong et al., 2022a) model, which is called **initial fine-tuned**. For SSDS, we set the selected ratio of unlabeled dialogues as 25% by default. The full unlabeled method uses all (100%) unlabeled dialogues. $k = 20, 8, 8$ for SAMSUM, DIALOGSUM, and TODSUM respectively. For uncertainty estimation, we rank all unlabeled dialogues and replace the most uncertain summaries with their ground truth summaries with a given ratio from [0%, 10%, ..., 90%]. We set the α, β, γ to 1 by default, where parameter analysis and search (based on ROUGE-1) are presented in Sec. 5.6.4. We abbreviate our searched m+BNN results as **m+BNN-s**. The repetitive experimental setting is detailed in Sec. A.3.4.

5.6 Experimental Results

5.6.1 Uncertainty Estimation Results

We present our uncertainty estimation results on SiCF scores in Table 2 and 16. In these tables, we only list the ROUGE-1 and BERTScore-F. The details of ROUGE-1, -2, -L, and BERTScore-F are drawn in Figure 4, 5, 6, 7, 8 and 9 in appendix. With these tables, we conclude as below.

SiCF score is an effective way to improve uncertainty estimation. In the two tables, the SiCF scores always outperform the random rank baseline. For example, in Table 2, all six settings indicate that the SiCF (m+BNN) improves at least 1 point BERTScore-F compared with random rank in both small and medium-size labeled settings. This shows that SiCF effectively quantifies the quality of generated summaries by semantic invariance, coverage, and faithfulness, surpassing the straightforward outcomes of the random baseline.

SiCF (m+BNN) performs better than SiCF (mean) and SiCF (BNN) in the vast settings. The difference between the three designs is slight, with less than 1 point divergence. However, SiCF (m+BNN) performs better than SiCF (mean) and SiCF (BNN) in the vast settings, except the ROUGE-1 on TODSUM (2:90) in Table 16. For BERTScore-F in Table 2, BNN performs better than m+BNN on DIALOGSUM, which verifies the effectiveness of our designed variant-length multi-label BNN. Nevertheless, the better performance of SiCF (m+BNN) in the other two datasets still indicates that a m+BNN is better in uncertainty estimation in general.

5.6.2 SSDS Results

Table 3, 8 list SSDS results. We conclude below.

SiCF score is effective to improve SSDS. On the two settings of SAMSUM, the three ways of SiCF are generally beneficial for the SSDS compared with random rank. Though the improvement is less than 1 point, our SiCF shows a much better SSDS improved ratio (198%) compared to random rank (112%) and full unlabeled in terms of improved ratio in BERTScore-F. For the medium-size labeled settings of DIALOGSUM, only SiCF (m+BNN) generally performs better than the random rank but SiCF (mean) and SiCF (BNN) do not. This indicates that the combination of mean and multi-label BNN is effective. We do not report SSDS results on DIALOGSUM 1:50, as its generated pseudolabels are too noisy for training, which is detailed in the caption of Table 5. As for the two settings of TODSUM, SiCF (BNN) generally improves at least 2 points in terms of ROUGE-1, 2 and L compared to random rank in the medium setting. These verify the effectiveness of our SiCF score in improving SSDS by selecting unlabeled dialogues via SiCF scores.

Our methods surpass pseudo oracle due to higher sample diversity. As a surprising finding, in Table 3, our SiCF (m+BNN) is higher than pseudo oracle in terms of ROUGE-1 (e.g., 45.14 VS. 44.92) and BERTScore-F in both SAMSUM 1:50 and DIALOGSUM 5:50 settings. The possible reason is that the initial fine-tuned model might be good at predicting a certain distribution of the unlabeled dialogues. As a result, selected unlabeled dialogues in the pseudo oracle may be less diverse than those from SiCF (m+BNN).

Using all the unlabeled dialogues is not the best choice because some samples have significant pseudolabel noise. Compared to the results of full unlabeled, all metrics generally indicate that selecting 25% high-quality unlabeled dialogues is better. This is because only a part of generated pseudolabels is beneficial to the SSDS learning, which is verified in Table 5. Thus, it is essential to select high-quality <dialogue, pseudolabel> pairs.

5.6.3 Ablation Studies

We list our ablation studies in Tables 4, 19, and 20 for SAMSUM datasets with 1:50 setting. In the tables, "SiCF" method uses all three components. "Only sein", "Only cov", and "Only fai" are methods only using respective component. The two table answer the below question.

Among the three components in the SiCF score, the coverage score performs better, while a combination of three parts achieves overall improvement. Concretely, based on the two tables, using three components together generally improves performance except in SiCF mean cases. Its possible reason is that the pretrained model used in faithfulness might not perform well in SiCF, as the hallucination in text summarization is still a challenging problem. But we believe future research could improve the hallucination detection. As a result, SiCF using BNN and m+BNN both have results when using all three components.

Table 4: Ablation study of three components in SAMSUM 1:50 in terms of ROUGE-1. We use ± to connect the mean and standard deviation among 4 times repetitive experiments with different random seeds. The full table is in Tables 19.

ROUGE-1	SiCF (m+BNN)	
	0-50% mean	0-90% mean
sein+cov+fai	**59.932±0.015**	**71.151±0.031**
only sein	59.533±0.011	70.675±0.010
only cov	59.811±0.010	71.022±0.005
only fai	59.104±0.004	70.037±0.004

5.6.4 Parameter Analysis & Human Eval

We conduct parameter analysis on TODSUM (2:90) with SiCF (m+BNN), where its SSDS result is in Table 21, and its uncertainty estimation results are in Table 22. Plus, the parameter search results are shown in Table 18. For more details, please refer to Sec. A.3.5. Human evaluation is in Sec. A.3.7.

6 Conclusion

To make use of unlabeled data and measure generated summary quality for summarization model training, we benchmark SSDS and uncertainty estimation on dialogue summarization. We propose the SiCF score, which measures semantic invariance, coverage, and faithfulness of pseudolabels (generated summaries) at the text level, word level, and sentence level, respectively. Furthermore, we extend BNN-based uncertainty estimation to a variant-length multi-label setting. Our SiCF score can enhance uncertainty estimation on dialogue summarization and improve SSDS by up to +1-2% ROUGE and BERTScore-F on SAMSUM (daily-chat domain), TODSUM (task-oriented dialogues) and DIALOGSUM (real-life scenario).

7 Ethical Consideration

This study pioneers the evaluation of summary quality without relying on ground truth summaries. During our study, we address the challenge of diverse ground truth summaries for dialogue summarization in an innovative way.

Our research employs datasets that are publicly available, ensuring transparency and accessibility. The datasets integral to our work are utilized in adherence to their respective licenses, which is verified in Sec. A.4.

All the datasets utilized in our study are devoid of personal identification details. We advise that any potential extensions of this research into domains containing personal or sensitive information should be conducted with strict adherence to robust ethical guidelines.

8 Limitations

This paper introduces the SiCF score as a means of evaluating the quality of generated summaries without using ground truth summaries. However, SiCF has a limitation: when we restrict the occurrence of proper nouns to a maximum of one, it affects other proper nouns that are not speaker names. Therefore, we need to explore alternative approaches for weighing each key detail and key point.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

Andy Chen. 2019. Is variance really a measure of uncertainty? `https://shorturl.at/ntN79`.

Jiaao Chen, Mohan Dodda, and Diyi Yang. 2022. Human-in-the-loop abstractive dialogue summarization. *arXiv preprint arXiv:2212.09750*.

Jiaao Chen and Diyi Yang. 2021. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616.

Wang Chen, Piji Li, Hou Pong Chan, and Irwin King. 2021a. Dialogue summarization with supporting utterance flow modelling and fact regularization. *Knowledge-Based Systems*, 229:107328.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021b. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Filipe R Cordeiro and Gustavo Carneiro. 2020. A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations? In *2020 33rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 9–16. IEEE.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2023. Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2044–2060.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, pages 1–77.

Alexios Gidiotis and Grigorios Tsoumakas. 2023. Bayesian active summarization. *Computer Speech & Language*, 83:101553.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Yile Gu, Prashanth Gurunath Shivakumar, Jari Kolehmainen, Ankur Gandhe, Ariya Rastrow, and Ivan Bulyko. 2023. Scaling laws for discriminative speech recognition rescoring models. *arXiv preprint arXiv:2306.15815*.

Som Gupta and Sanjai Kumar Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Jianfeng He, Julian Salazar, Kaisheng Yao, Haoqi Li, and Jinglun Cai. 2023a. Zero-shot end-to-end spoken language understanding via cross-modal selective self-training. *arXiv preprint arXiv:2305.12793*.

Jianfeng He, Linlin Yu, Shuo Lei, Chang-Tien Lu, and Feng Chen. 2023b. Uncertainty estimation on sequential labeling via uncertainty transmission. *arXiv preprint arXiv:2311.08726*.

Jianfeng He, Xuchao Zhang, Shuo Lei, Abdulaziz Alhamadani, Fanglan Chen, Bei Xiao, and Chang-Tien Lu. 2023c. Clur: Uncertainty estimation for few-shot text classification with contrastive learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 698–710.

Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and ChangTien Lu. 2020. Towards more accurate uncertainty estimation in text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8362–8372.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788*.

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. *arXiv preprint arXiv:2301.10483*.

Wojciech Kryscinski and Bryan McCann. 2021. Evaluating the factual consistency of abstractive text summarization. US Patent App. 16/750,598.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

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.

Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, and Chang-Tien Lu. 2022. Uncertainty-aware cross-lingual transfer with pseudo partial labels. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1987–1997.

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

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

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

Yang Liu, Sheng Shen, and Mirella Lapata. 2020. Noisy self-knowledge distillation for text summarization. *arXiv preprint arXiv:2009.07032*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu and Nancy F Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. *arXiv preprint arXiv:2109.13070*.

Nishant Mishra, Gaurav Sahu, Iacer Calixto, Ameen Abu-Hanna, and Issam H Laradji. 2023. Llm aided semi-supervision for extractive dialog summarization. *arXiv preprint arXiv:2311.11462*.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. 2023. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394.

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*.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Gaurav Sahu, Olga Vechtomova, and Issam H Laradji. 2023. Enchancing semi-supervised learning for extractive summarization with an llm-based pseudolabeler. *arXiv preprint arXiv:2311.09559*.

Akim Tsvigun, Ivan Lysenko, Danila Sedashov, Ivan Lazichny, Eldar Damirov, Vladimir Karlov, Artemy Belousov, Leonid Sanochkin, Maxim Panov, Alexander Panchenko, et al. 2023. Active learning for abstractive text summarization. *arXiv preprint arXiv:2301.03252*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan O Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. *arXiv preprint arXiv:2305.14106*.

Mingye Wang, Pan Xie, Yao Du, and Xiaohui Hu. 2023. T5-based model for abstractive summarization: A semi-supervised learning approach with consistency loss functions. *Applied Sciences*, 13(12):7111.

Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)*, pages 985–992.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. *arXiv preprint arXiv:2106.09449*.

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of llms in implicit hate speech detection. *arXiv preprint arXiv:2402.11406*.

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

Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019b. Mitigating uncertainty in document classification. *arXiv preprint arXiv:1907.07590*.

Lulu Zhao, Fujia Zheng, Keqing He, Weihao Zeng, Yuejie Lei, Huixing Jiang, Wei Wu, Weiran Xu, Jun Guo, and Fanyu Meng. 2021. Todsum: Task-oriented dialogue summarization with state tracking. *arXiv preprint arXiv:2110.12680*.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022a. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

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. *arXiv preprint arXiv:2210.07197*.

Yingying Zhuang, Jiecheng Song, Narayanan Sadagopan, and Anurag Beniwal. 2023. Self-supervised pre-training and semi-supervised learning for extractive dialog summarization. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1069–1076.

A Appendix

A.1 More related work

Semi-supervised text summarization. There are two general types of text summarization: extractive text summarization (Wong et al., 2008; Liu, 2019), which extracts the original sentences from the text for summarization, and abstractive text summarization (Liu and Lapata, 2019; Nallapati et al., 2016), which directly generates summaries from the given text. For semi-supervised text summarization, Sahu et al. (2023) acknowledge that it is heavily under-explored.

In terms of semi-supervised extractive text summarization, Sahu et al. (2023) use a prompt-based method to implement extractive summarization, but their prompt is based on a subset of extracted sentences rather than the full text. In contrast, our work considers the full text.

In terms of abstractive text summarization, Wang et al. (2023) employ transfer learning for semi-supervised abstractive text summarization instead of quality assessment. While Gidiotis and Tsoumakas (2023) assess the quality of the generated summary, they measure the disagreement between each pair of generated summaries for a text via Monte Carlo (MC) dropout (Gal and Ghahramani, 2016). In contrast, our semantic invariance has a lower time complexity due to no comparison among each pair. Additionally, our coverage and faithfulness consider the relation between dialogues and generated summaries, while Gidiotis and Tsoumakas (2023) do not.

Also related to unlabeled text, Tsvigun et al. (2023) focus on active learning via the similarity between labeled and unlabeled text. In contrast, we focus on semi-supervised learning, and our SiCF score considers the relation between pseudo labels and unlabeled dialogues.

Summary quality. Summary quality measurement can be divided into: with ground truth and without ground truth. Previous quality measurement methods mostly rely on the availability of ground truth, such as ROUGE Scores (Lin, 2004) and BERT Scores (Zhang et al., 2019a). Additionally, criteria like coverage and faithfulness were proposed as metrics to evaluate summarization in Huang et al. (2023), which also uses ground truth summary. Faithfulness can also be evaluated Dreyer et al. (2023) through automated evaluation, as proposed in FactCC (Kryscinski and McCann, 2021) and DOCNLI (Yin et al., 2021). Besides, coherence is proposed in Zhong et al. (2022b). However, these evaluations all require ground truth summaries. In contrast, the SSDS task is inaccessible to ground truth, and the measurement of summary quality without relying on ground truth is underexplored. Although semantic uncertainty was introduced to assess generation quality from the view of uncertainty (Kuhn et al., 2023), it is time-consuming and overlooks the relationship between the generation and context. Consequently, we propose a more efficient method for measuring semantic invariance. Moreover, we are the first to comprehensively evaluate summary quality focusing on semantic invariance, coverage, and faithfulness, without relying on ground truth summaries.

A.2 Model

A.2.1 Reasons for choosing semantic invariance, coverage, and faithfulness

Inspired by Huang et al. (2023), our justification for choosing semantic invariance, coverage, and faithfulness is based on three key requirements. Firstly, we aim to perform quality analysis within the generated summaries (semantic invariance) and evaluate the quality between the generated summaries and the quality analysis (coverage and faithfulness). Secondly, we aim to conduct quality evaluations at the token level (coverage), sentence level (faithfulness), and text level (semantic invariance). Thirdly, we strive for our quality evaluation to be inaccessible to ground truth summaries, which are often unavailable in real-world scenarios. Our designed three components fulfill these three requirements.

Additionally, our coverage and faithfulness evaluations do not rely on ground truth summaries, whereas the coverage and faithfulness evaluations in Huang et al. (2023) do.

A.2.2 Comparison between POS and NER model

To extract key information from the dialogues, the intuitive choice is the Named Entity Recognition (NER) model, while we find that most of the extracted entities from 'ner-english-ontonotes' are all speaker names. In contrast, we find that nouns can represent the key information of the dialogues in a good way.

An example of comparing the use of a NER model and a Part-of-Speech (POS) model to extract key information is shown below, where the blue font indicates the annotated key information.

Table 5: SSDS results on TODSUM 2:90 and 10:90 settings with 50% ratio. From the table, we can see that the pseudo oracle using 25% unlabeled data performs better than the pseudo oracle using 50% unlabeled data, and even better than that using full (100%) unlabeled data. Due to the high level of noise in the pseudolabels of the DIALOGSUM dataset in the small-sized labeled setting, the pseudo oracle compared to the initially fine-tuned model shows only marginal improvement.

	Select Ratio	1:50 or 2:90				5:50 or 10:90			
		ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F
SAMSUM									
Initial fine-tuned	N/A	43.90	18.49	35.02	0.4374	46.81	20.46	36.38	45.74
Full Unlabeled	100%	44.32	19.07	35.43	43.72	47.76	21.84	**38.20**	47.02
Pseudo Oracle	50%	**45.58**	19.79	**35.95**	44.00	**48.40**	21.73	38.16	**47.57**
Pseudo Oracle	25%	44.92	**19.87**	35.67	**44.32**	48.33	**21.97**	37.93	47.32
DIALOGSUM									
Initial Fine-Tuned	N/A	40.30	14.53	31.19	43.89	42.28	15.61	33.18	46.29
Full Unlabeled	100%	39.22	14.04	30.50	43.49	41.61	16.04	33.25	46.12
Pseudo Oracle	50%	40.37	14.54	31.44	44.69	42.44	**16.37**	**33.80**	46.87
Pseudo Oracle	25%	**40.44**	**14.98**	**31.73**	**45.43**	**42.71**	16.10	33.77	**47.05**
TODSUM									
Initial Fine-Tuned	N/A	76.60	59.51	67.87	70.02	79.75	64.69	72.77	74.28
Full Unlabeled	100%	77.02	59.86	68.09	70.57	80.00	64.99	73.08	74.30
Pseudo Oracle	50%	78.44	61.65	70.11	72.37	82.49	68.48	76.40	77.18
Pseudo Oracle	25%	**79.09**	**63.19**	**71.94**	**73.96**	**83.43**	**70.10**	**77.43**	**78.48**

1. Dialogue: "Tom: Happy B-day! Tom: <file gif> Laura: oh , thank you , it's so cute <3 <3 <3 Tom: :D"
2. Ground truth summaries: Tom wishes Laura happy brithday.
3. Predicted summaries:Laura thanked Tom for his birthday.

The NER model extracted results are shown as below:
1. Dialogue: {Tom: Person, Tom: Person, Laura: Person}
2. Ground truth summaries: {Tom: Person, Laura: Person}
3. Predicted summaries: {Tom: Person, Laura: Person}

Here, we observe that the extracted entities are all person names. Given that the dialogues often contain many speaker names, which cannot effectively differentiate between different dialogues, we believe that using a NER model to extract key information may not be ideal.

The POS model extracted results are shown below, where NNP represents Proper noun, singular, and NN represents Noun, singular or mass.
1. Dialogue: {Tom: NNP, B-day: NN, Tom: NNP, file: NN, Laura: NNP, Tom: NN, D: NN}
2. Ground truth summaries: {Tom: NNP, Laura: NNP, B-day: NN}
3. Predicted summaries: {Tom: NNP, Laura: NNP, B-day: NN}

In this case, we find that the POS results for the dialogue have a large overlap with the key information from a human perspective (the blue fonts in the original dialogue). Thus, the nouns in the dialogue are a better representation of key information compared to the NER model.

A.2.3 Multi-label BNN

$$\lambda_{cov/fai} = \underbrace{\sum_{l=1}^{p/h} \mathbb{H}[Y_l|x, D]}_{Predictive} = \underbrace{\sum_{l=1}^{p/h} \mathbb{I}[Y_l|x, D]}_{Epistemic} + \underbrace{\sum_{l=1}^{p/h} \mathbb{E}[\mathbb{H}[Y_l|x, D]]}_{Aleatoric} \qquad (10)$$

In BNN theory (Gawlikowski et al., 2023), predictive uncertainty usually consists of aleatoric uncertainty and epistemic uncertainty.

The aleatoric uncertainty is irreducible because it refers to the noise in data generation, such as imperfect sensors. And the epistemic uncertainty refers to a model uncertainty due to limited knowledge, such as having insufficient training data. The predictive BNN is the sum of these two items.

To get the three kinds of uncertainty, we usually firstly calculate the predictive uncertainty as the mean $\bar{Y}_l$ of $Y_l \in \mathbb{R}^{k\times 2}$ for l-th label, followed by calculating the entropy of $\bar{Y}_l$. Finally, the sum of all labels' entropy is the predictive uncertainty score in terms of coverage λ_{cov} and faithfulness score λ_{fai}. To obtain aleatoric uncertainty, we calculate the entropy of each Y_l at first, before calculating the expectation of all k entropy values; finally, the sum of all expectations of l labels' entropy expectation is the aleatoric uncertainty score. As for epistemic uncertainty, it is usually obtained by using a predictive uncertainty score to subtract its epistemic uncertainty.

A.2.4 Special Case on Faithfulness

To benefit the understanding of faithfulness, we omit a special case for faithfulness in Sec. 4.4, where a dialogue sentence has no nouns. Concretely, when i-th dialogue sentence in a dialogue has no nouns, the i-th element in $\tilde{v}^b \in \mathbb{R}^h = min(\tilde{V}^b)$ will equal 0, because $w_i^b = 0$ in Eq. 4. We do not expect this to happen for a dialogue sentence without nouns, as a smaller value in $\tilde{v}^b$ refers to better faithfulness. Therefore, we add an activation A_{w^b} on $\tilde{v}^b$, followed by concatenating k activated results of $A_{w^b}(\tilde{v}^b)$ to obtain $\tilde{V}^b$. $A_{w^b}(\tilde{v}^b)$ is formulated as below,

$$A_{w_i^b}(\tilde{v}_i^b) = \begin{cases} \tilde{v}_i^b, & w_i^b \geq 0 \\ \varpi, & w_i^b = 0 \end{cases} \tag{11}$$

where it keeps the original $\tilde{v}_i^b$ if the respective i-th sentence in a dialogue has at least one noun, or else gives a large scalar ϖ.

A.3 More Experimental Results

A.3.1 More Explanation About SSDS Improved Ratio

If a method's improved ratio is greater than 100%, it indicates superior performance compared to the pseudo oracle. Conversely, if a method's improved ratio is smaller than 0, it suggests worse performance than the initial dialogue summarization that is fine-tuned only on the labeled samples. A higher improved ratio for a method signifies its superiority over another method.

A.3.2 Uncertainty Estimation Results

For the aleatoric uncertainty is more important in uncertainty estimation than epistemic uncertainty. In the uncertainty estimation task, based on Table 10 and 11, we see that aleatoric performs better than epistemic in both SiCF (BNN) and SiCF (m+BNN). This indicates that the aleatoric uncertainty (such as noise in the sample collection impacts more than the epistemic uncertainty (such as insufficient training samples). It further shows that pseudolabel noise impacts more compared with the unlabeled sample size in SSDS on the SAMSUM 1:50 setting. Also, in SiCF (BNN), using both aleatoric and epistemic (our default usage) improves the uncertainty estimation. In contrast, in SiCF (m+BNN), only using both aleatoric (our default usage) improves the uncertainty estimation. The possible reason is that the mean information is more complementary to aleatoric compared with epistemic. However, in our experiments, we still use the predictive uncertainty, a sum of aleatoric and epistemic uncertainty, which is a commonly usage of the two kinds of uncertainty.

Besides the listed mean values of 0-50% and 0-90% in the Table 16 and 2, we also draw the concrete metric values in different force true ratios in Figure 4, 5, 6, 7, 8, 9.

A.3.3 SSDS Results

BNN and m+BNN perform better than mean in SSDS. Based on Table 12, we found among the 12 comparisons, two groups show mean performs better, five groups show that BNN performs better, and five groups show that m+BNN performs better. This demonstrates the positive effect of BNN and a combination of the mean and BNN on SSDS.

Epistemic uncertainty benefits SSDS performance than aleatoric uncertainty. From Table 17, we see that using both aleatoric and epistemic uncertainty generally leads to better SSDS results. For example, Considering Tables 10 and 11, we conclude that though aleatoric uncertainty benefits in improving uncertainty estimation results, using both aleatoric and epistemic uncertainty benefits in improving SSDS results.

A.3.4 Robust Experimental Settings

To assess the robustness of our experiments, we conducted our experiments four times, in addition to the original single run. Each of these four runs used different random seeds, but they all shared the same initial fine-tuned model and used the same set of beam search sampling to generate summaries. This approach allows us to evaluate the consistency of our SiCF scores in a semi-supervised setting, where the initial fine-tuned model is not our focus and should be consistent for fairness. Furthermore, these repeated experiments employed the same set of beam search sampling generated summaries for a fair assessment of robustness. For example, in the SAMSUM 1:50 setting, all four experiments were provided with the same set of 20 generated summaries for each unlabeled dialogue. The results of these robustness experiments are presented in Tables 4, 19, 13, 14, and 15. Obtaining Table 15 with an additional 3 rounds of experiments for all 7 methods requires approximately 252 hours (10.5 days) on a server equipped with 4 V100 GPUs.

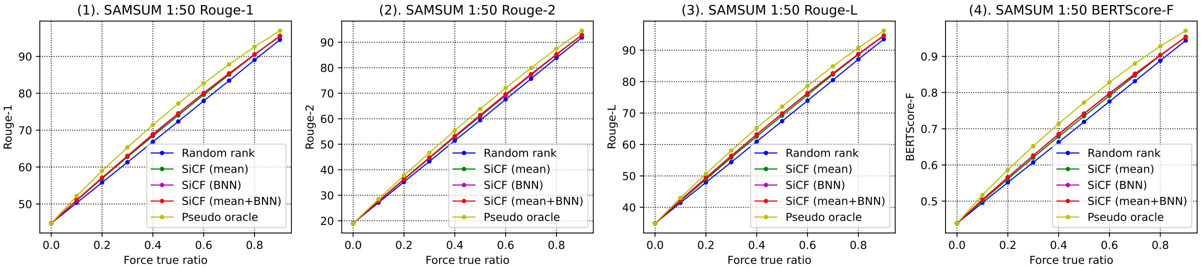


Figure 4: Diagram of uncertainty estimation results in force true ratio of 0%, 10%, 20% ..., 90% on SAMSUM 1:50 setting.

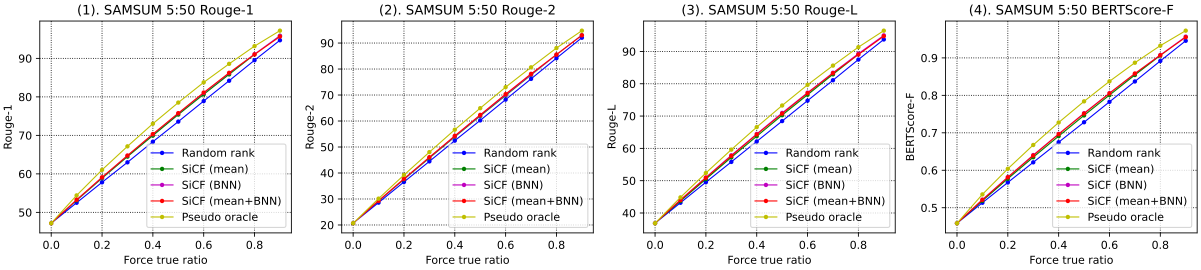


Figure 5: Diagram of uncertainty estimation results in force true ratio of 0%, 10%, 20% ..., 90% on SAMSUM 5:50 setting.

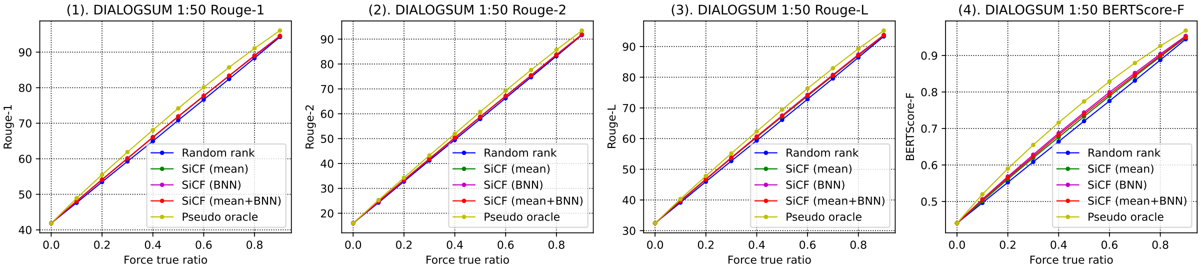


Figure 6: Diagram of uncertainty estimation results in force true ratio of 0%, 10%, 20% ..., 90% on DIALOGSUM 1:50 setting.

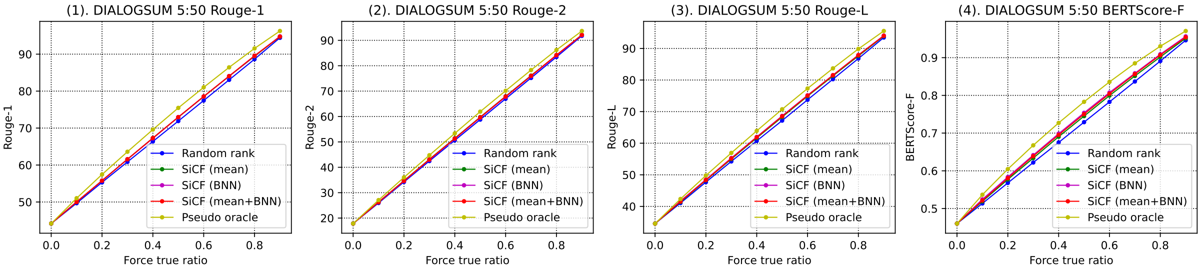


Figure 7: Diagram of uncertainty estimation results in force true ratio of 0%, 10%, 20% ..., 90% on DIALOGSUM 5:50 setting.

Table 6: SSDS results on SAMSUM 1:50 setting and on TODSUM 2:90 setting. The values in the bracket are SSDS improved ratios.

	1:50 or 2:90			
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F
SAMSUM				
Initial Fine-Tuned	43.90(0%)	18.49(0%)	35.02(0%)	43.74(0%)
Full Unlabeled	44.32(41%)	19.07(42%)	35.43(63%)	43.72(-3%)
Random Rank	44.98(105%)	19.32(60%)	35.65(96%)	44.39(112%)
SiCF (mean)	**45.85(191%)**	19.90(102%)	**35.96(144%)**	**44.89(198%)**
SiCF (BNN)	45.20(127%)	**19.95(105%)**	35.63(93%)	44.45(122%)
SiCF (m+BNN)	45.14(121%)	19.31(59%)	35.59(87%)	44.47(125%)
SiCF (m+BNN-s)	45.40(147%)	19.38(64%)	35.48(70%)	44.34(103%)
Pseudo Oracle	44.92(100%)	19.87(100%)	35.67(100%)	44.32(100%)
TODSUM				
Initial Fine-Tuned	76.60(0%)	59.51(0%)	67.87(0%)	70.02(0%)
Full Unlabeled	77.02(16%)	59.86(9%)	68.09(5%)	70.57(13%)
Random Rank	76.91(12%)	59.43(-2%)	68.29(10%)	70.64(15%)
SiCF (mean)	**77.94(53%)**	**61.01(40%)**	**69.64(43%)**	**71.72(43%)**
SiCF (BNN)	76.64(1%)	59.66(4%)	68.71(20%)	70.53(12%)
SiCF (m+BNN)	77.01(16%)	59.92(11%)	69.01(28%)	70.91(22%)
SiCF (m+BNN-s)	76.45(-6%)	59.46(-1%)	68.77(22%)	70.64(15%)
Pseudo Oracle	79.09(100%)	63.19(100%)	71.94(100%)	73.96(100%)

Table 7: SSDS results on SAMSUM 5:50 setting and on TODSUM 10:90 setting. The values in the bracket are SSDS improved ratios.

	5:50 or 10:90			
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F
SAMSUM				
Initial Fine-Tuned	46.81(0%)	20.46(0%)	36.38(0%)	45.74(0%)
Full Unlabeled	47.76(62%)	21.84(91%)	**38.20(117%)**	**47.02(81%)**
Random Rank	47.65(55%)	21.13(44%)	37.18(51%)	46.66(58%)
SiCF (mean)	47.90(71%)	21.39(61%)	37.57(76%)	46.51(48%)
SiCF (BNN)	47.77(63%)	**22.07(106%)**	37.67(83%)	46.35(38%)
SiCF (m+BNN)	**48.14(87%)**	22.05(105%)	37.63(80%)	46.61(55%)
SiCF (m+BNN-s)	47.83(67%)	21.40(62%)	37.24(55%)	46.55(51%)
Pseudo Oracle	48.33(100%)	21.97(100%)	37.93(100%)	47.32(100%)
TODSUM				
Initial Fine-Tuned	79.75(0%)	64.69(0%)	72.77(0%)	74.28(0%)
Full Unlabeled	80.00(6%)	64.99(5%)	73.08(6%)	74.30(0%)
Random Rank	80.57(22%)	65.73(19%)	73.76(21%)	75.13(20%)
SiCF (mean)	81.08(36%)	66.70(37%)	74.48(36%)	75.75(34%)
SiCF (BNN)	**82.01(61%)**	**67.90(59%)**	**75.95(68%)**	**76.74(58%)**
SiCF (m+BNN)	80.93(32%)	66.32(30%)	74.58(38%)	75.78(35%)
SiCF (m+BNN-s)	81.22(39%)	67.14(45%)	75.16(51%)	76.15(44%)
Pseudo Oracle	83.43(100%)	70.10(100%)	77.43(100%)	78.48(100%)

Table 8: SSDS results on DIALOGSUM 5:50 settings. We do not report SSDS results on DIALOGSUM 1:50, as its generated pseudolabels are too noisy for training and detailed in the caption of Tab. 5.

	Medium-size Labeled Data			
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F
DIALOGSUM				
Initial Fine-Tuned	42.28(0%)	15.61(0%)	33.18(0%)	46.29(0%)
Full Unlabeled	41.61(-155%)	16.04(87%)	33.25(11%)	46.12(-22%)
Random Rank	42.32(9%)	16.38(157%)	33.70(88%)	46.07(-28%)
SiCF (mean)	41.77(-118%)	15.87(53%)	32.79(-66%)	45.48(-106%)
SiCF (BNN)	42.00(-65%)	15.95(69%)	33.47(49%)	46.70(53%)
SiCF (m+BNN)	42.85(132%)	16.86(255%)	34.06(149%)	46.45(21%)
SiCF (m+BNN-s)	**43.02(172%)**	**17.22(328%)**	**34.32(193%)**	**47.02(96%)**
Pseudo Oracle	42.71(100%)	16.10(100%)	33.77(100%)	47.05(100%)

A.3.5 Parameter Analysis & Search

We conduct a parameter analysis on TODSUM 2:90 with SiCF (m+BNN), where its SSDS result

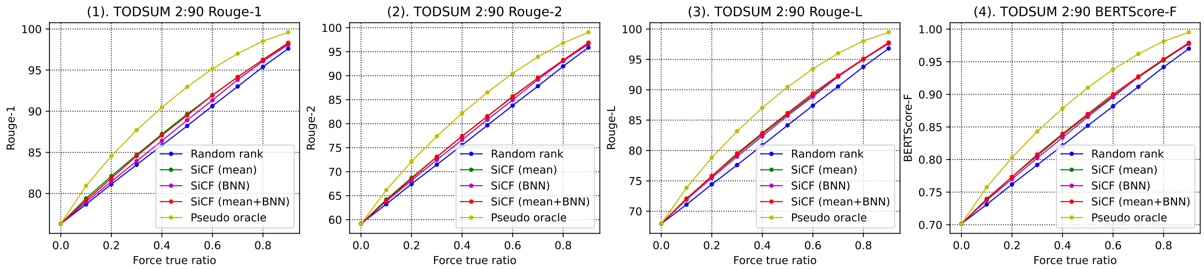


Figure 8: Diagram of uncertainty estimation results in force true ratio of 0%, 10%, 20% ..., 90% on TODSUM 2:90 setting.

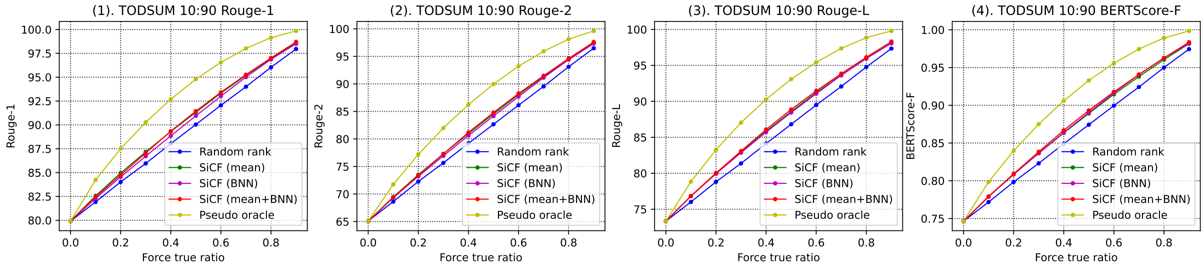
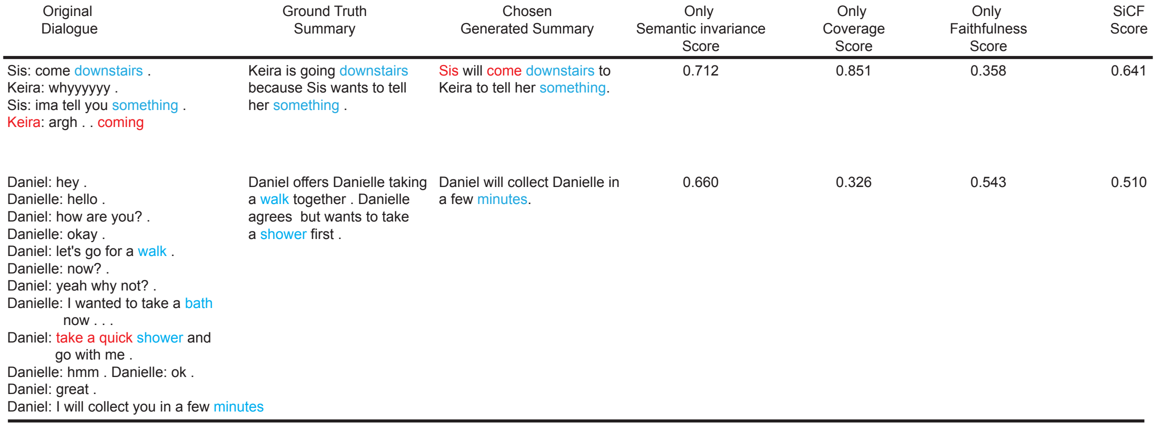


Figure 9: Diagram of uncertainty estimation results in force true ratio of 0%, 10%, 20% ..., 90% on TODSUM 10:90 setting.



Original Dialogue	Ground Truth Summary	Chosen Generated Summary	Only Semantic invariance Score	Only Coverage Score	Only Faithfulness Score	SiCF Score
Sis: come downstairs . Keira: whyyyyyy . Sis: ima tell you something . Keira: argh . . coming	Keira is going downstairs because Sis wants to tell her something .	Sis will come downstairs to Keira to tell her something.	0.712	0.851	0.358	0.641
Daniel: hey . Danielle: hello . Daniel: how are you? . Danielle: okay . Daniel: let's go for a walk . Danielle: now? . Daniel: yeah why not? . Danielle: I wanted to take a bath now . . . Daniel: take a quick shower and go with me . Danielle: hmm . Danielle: ok . Daniel: great . Daniel: I will collect you in a few minutes	Daniel offers Danielle taking a walk together . Danielle agrees but wants to take a shower first .	Daniel will collect Danielle in a few minutes.	0.660	0.326	0.543	0.510

Figure 10: The quality analysis of our SiCF score as well as its ablation study. Regarding the semantic invariance score, the k diverse generated summaries are not displayed for brevity. In the coverage score, key details (nouns) are highlighted in blue. For assessing faithfulness, conflicts (first row) and missing information (second row) are indicated in red. The analysis of the figure is in Sec. A.3.6. Please zoom in for better visualization.

Table 9: Human evaluation results on the TODSUM on 100 testing samples from the SSDS task in the medium-size setting, assessed by 5 participants.

Method	Human preference rate ↑
Full Unlabeled	21.20%
SiCF(mean)	23.40%
SiCF(BNN)	22.80%
SiCF(m+BNN)	**32.60%**

comparison on 25% ration is shown in Table 21, and its uncertainty estimation results are presented in Table 22.

For this section, we can conclude as below.

The Impact of Parameters on Uncertainty Estimation and SSDS. Based on Table 22, we observe that enlarging the coefficient of semantic invariance can improve uncertainty estimation results by 0-50% on TODSUM 2:90. However, the 0-90% mean performs better when the coefficients of the three components are balanced.

Additionally, according to Table 21, the best SSDS performance is achieved when the coefficients of the three components are balanced. This suggests that each component in SiCF scores con-

Table 10: Comparison of aleatoric and epistemic uncertainty estimation results on SAMSUM 1:50 in terms of ROURE-1 and ROUGE-2.

	ROUGE-1		ROUGE-2	
	0-50% mean	0-90% mean	0-50% mean	0-90% mean
SiCF(BNN, alea+epis)	**59.84**	**71.03**	**40.34**	**56.66**
SiCF(BNN, alea)	59.80	70.98	40.30	56.58
SiCF(BNN, epis)	59.70	70.85	40.21	56.53
SiCF(m+BNN, alea+epis)	59.91	71.10	40.41	56.77
SiCF(m+BNN, alea)	**59.94**	**71.17**	**40.41**	**56.78**
SiCF(m+BNN, epis)	59.83	70.99	40.35	56.67

Table 11: Comparison of aleatoric and epistemic uncertainty estimation results on SAMSUM 1:50 in terms of ROUGE-L and BERTScore-F.

	ROUGE-L		BERTScore-F	
	0-50% mean	0-90% mean	0-50% mean	0-90% mean
SiCF(BNN, alea+epis)	**52.69**	**65.87**	59.34	**70.64**
SiCF(BNN, alea)	52.68	65.83	**59.35**	**70.64**
SiCF(BNN, epis)	52.42	65.58	59.09	70.40
SiCF(m+BNN, alea+epis)	52.62	65.80	59.38	70.69
SiCF(m+BNN, alea)	**52.68**	**65.89**	**59.45**	**70.78**
SiCF(m+BNN, epis)	52.53	65.69	59.24	70.53

Table 12: Comparison of mean (m), BNN (B), and m+BNN (m+B) on SSDS results based on Table 3. ">" means better.

	SAMSUM		DIALOGSUM		TODSUM	
	1:50	5:50	1:50	5:50	2:90	10:90
50%	B>m+B>m	m+B>B>m	B>m+B>m	m+B>B>m	m+B>m>B	B>m+B>m
25%	m>B>m+B	B>m+B>m	m+B>B>m	m+B>B>m	m>m+B>B	B>m+B>m

Table 13: Mean and standard deviation of uncertainty estimation results of SiCF score on TODSUM datasets in terms of ROUGE-1 on the medium-size labeled data. The results are reported based on four repetitive experiments via different random seeds. The "0-50" and "0-90" are the mean of ROUGE-1 with eliminated ratio range 0%-50% and 0%-90%. It is respective to the Table 16.

ROUGE-1	TODSUM (10:90)	
	0-50%	0-90%
Random Rank	84.963±0.027	88.980±0.042
SiCF(mean)	85.822±0.028	89.823±0.056
SiCF(BNN)	85.509±0.024	89.629±0.024
SiCF(m+BNN)	85.774±0.033	89.887±0.031
SiCF(m+BNN-s)	**85.836±0.004**	**89.949±0.002**
Pseudo Oracle	88.235±0.003	92.298±0.001

tributes to the quality measurement of pseudolabels.

The best-searched coefficient on the uncertainty estimation. We listed the best-searched coefficient in Table 18 by searching the coefficient leading to the best ROUGE-1 on the uncertainty estimation task. The search range is [0, 0.25, 0.5, 0.75, 1] for each of the three component coefficients. Based on the table, there is no 0 coefficient. This also indicates that each component benefits the uncertainty estimations in all six settings.

Table 14: Mean and standard deviation of uncertainty estimation results of SiCF score on TODSUM datasets in terms of BERTScore-F on the medium-size labeled data. The results are reported based on four repetitive experiments via different random seeds. The "0-50" and "0-90" are the mean of BERTScore-F with eliminated ratio range 0%-50% and 0%-90%. It is respective to the Table 2.

BERTScore-F	TODSUM (10:90)	
	0-50%	0-90%
Random Rank	81.02±0.017	86.086±0.032
SiCF(mean)	81.935±0.067	87.016±0.089
SiCF(BNN)	82.035±0.026	87.221±0.028
SiCF(m+BNN)	82.198±0.019	87.355±0.014
SiCF(m+BNN-s)	**82.273±0.002**	**87.433±0.002**
Pseudo Oracle	84.977±0.002	90.163±0.002

A.3.6 Quality Analysis of SSDS Results

Figure 10 presents the quality analysis of SSDS results. Specifically, we observe that the chosen generated summaries in the first row have higher coverage, including terms like "downstairs" and "something." However, they differ from the original dialogue in terms of who comes downstairs, resulting in a high coverage score but a low faithfulness score.

In contrast, the generated summaries in the

Table 15: Mean and standard deviation of SSDS results on TODSUM medium-size labeled data. The results are reported based on four repetitive experiments via different random seeds. It is respective to Table 3.

	Medium-Size Labeled Data		
	ROUGE-1	ROUGE-2	BERTScore-F
	TODSUM		
Full Unlabeled	80.19±0.33	65.34±0.63	74.57±0.41
Random Rank	80.66±0.14	66.06±0.39	75.19±0.2
SiCF(mean)	81.49±0.26	67.13±0.33	76.06±0.22
SiCF(BNN)	**82.28±0.18**	**68.40±0.32**	**77.04±0.19**
SiCF(m+BNN)	81.72±0.47	67.47±0.67	76.50±0.44
SiCF(m+BNN-s)	81.29±0.11	67.01±0.17	76.05±0.11
Pseudo Oracle	83.75±0.25	70.52±0.32	78.72±0.19

Table 16: Uncertainty estimation results of SiCF score on three datasets in terms of ROUGE-1. The “0-50” and “0-90” are the mean of ROUGE-1 scores with eliminated ratio range 0%-50% and 0%-90%. In the medium-size setting, the TODSUM has a mean and standard deviation as shown in Table 13.

ROUGE-1	SAMSUM(1:50)		DIALOGSUM(1:50)		TODSUM (2:90)		SAMSUM(5:50)		DIALOGSUM(5:50)		TODSUM(10:90)	
	0-50%	0-90%	0-50%	0-90%	0-50%	0-90%	0-50%	0-90%	0-50%	0-90%	0-50%	0-90%
Random Rank	58.55	69.62	56.34	67.97	82.28	87.04	60.40	70.97	58.05	69.18	84.98	88.99
SiCF(mean)	59.61	70.83	57.02	68.65	83.24	88.00	61.50	72.24	58.63	69.85	85.87	89.92
SiCF(BNN)	59.84	71.03	57.01	68.68	82.67	87.54	61.71	72.44	58.61	69.86	85.55	89.67
SiCF(m+BNN)	59.91	71.10	57.03	68.67	83.08	87.92	61.74	72.45	58.64	69.88	85.83	89.94
SiCF(m+BNN-s)	**59.98**	**71.20**	**57.06**	**68.75**	**83.19**	**87.99**	**61.86**	**72.61**	**58.70**	**69.96**	**85.84**	**89.95**
Pseudo Oracle	61.63	72.98	58.41	70.35	85.49	90.32	63.56	74.41	60.22	71.66	88.23	92.30

Table 17: Comparison of aleatoric and epistemic of SSDS results on SAMSUM 1:50

	Select Ratio	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F
SiCF(BNN, alea+epis)	0.25	**45.20**	**19.95**	35.63	**44.45**
SiCF(BNN, alea)	0.25	44.49	18.93	34.77	43.51
SiCF(BNN, epis)	0.25	45.35	19.69	**35.77**	44.39
SiCF(m+BNN, alea+epis)	0.25	**45.14**	19.31	**35.59**	**44.47**
SiCF(m+BNN, alea)	0.25	45.06	19.02	35.19	44.16
SiCF(m+BNN, epis)	0.25	45.07	**19.35**	35.49	44.19

Table 18: The coefficient of hyperparameter search in terms of ROUGE-1. We use “sein” to represent semantic invariance, apply “cov” to represent coverage and utilize “fai” to denote faithfulness.

	SAMSUM		DIALOGSUM		TODSUM	
	1:50	5:50	1:50	5:50	2:90	10:90
SeIn	0.5	0.5	1	1	0.75	0.75
Cov	0.75	1	1	0.5	0.25	0.5
Fai	0.25	0.25	0.25	0.25	0.5	0.5

second row lack terms like “walk,” “bath,” and “shower,” leading to a low coverage score. Although they do not have a conflict similar to the first row, they miss the shower-related content, resulting in relatively low faithfulness scores.

A.3.7 Human Evaluation Results

We had 5 participants for the human evaluation test. We presented each person with the same 100 testing samples, including the original dialogues, ground truth dialogue summaries, and four randomly-ordered generated dialogue summaries from four methods (Full Unlabeled, SiCF (mean), SiCF (BNN), SiCF (mean+BNN)). Subsequently, each person was asked the question “Among the four generated dialogue summaries, which is the best when compared to the original dialogues and ground truth dialogue summaries?”

From Table 9, we can see that SiCF (mean + BNN) performs the best from the human perspective with a 32.60% human preference rate. The human preference rate is the ratio between the total respective selected samples and the total 500 times selection.

A.3.8 More Experimental Settings

Our experiments run on 4 V100 GPUs, with 12 hours on SAMSUM for the full training.

A.4 License Analysis

The SAMSUM dataset is licensed under CC BY-NC-ND 4.0. The DIALOGSUM dataset is licensed under the MIT License. As for TODSUM, it is publicly released without a specified license. Therefore, our research usage of these datasets complies with their respective licenses.

Table 19: Ablation study of three components in SAMSUM 1:50 in terms of ROUGE-1. We use “sein” to represent semantic invariance, apply “cov” to represent coverage, and utilize “fai” to denote faithfulness. We use ± to connect the mean and standard deviation among 4 times repetitive experiments with different random seeds. A standard deviation of 0.000 means that differences occur in the decimal places further to the right.

ROUGE-1	SiCF (Mean)		SiCF (BNN)		SiCF (m+BNN)	
	0-50% mean	0-90% mean	0-50% mean	0-90% mean	0-50% mean	0-90% mean
sein+cov+fai	59.941±0.191	71.139±0.177	**59.812±0.016**	**70.996±0.021**	**59.932±0.015**	**71.151±0.031**
only sein	59.533±0.011	70.675±0.010	59.533±0.011	70.675±0.010	59.533±0.011	70.675±0.010
only cov	**59.964±0.006**	**71.183±0.005**	59.676±0.005	70.812±0.005	59.811±0.010	71.022±0.005
only fai	58.657±0.026	69.794±0.037	59.553±0.006	70.671±0.005	59.104±0.004	70.037±0.004

Table 20: Ablation study of three components in SAMSUM 1:50 in terms of BERTScore F. The organization is similar to Table 19.

BERTScore-F	SiCF (Mean)		SiCF (BNN)		SiCF (m+BNN)	
	0-50% mean	0-90% mean	0-50% mean	0-90% mean	0-50% mean	0-90% mean
sein+cov+fai	58.90	70.28	**59.34**	**70.64**	**59.38**	**70.69**
only sein	58.83	70.14	58.83	70.14	58.83	70.14
only cov	**59.21**	**70.56**	59.26	70.52	5932	70.63
only fai	58.01	69.32	59.06	70.31	58.61	69.69

Table 21: Parameter analysis of SSDS results on TODSUM 2:90, where the SiCF scores are all calculated by m+BNN. The values in the brackets are α, β, and γ, respectively.

	Select Ratio	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F
SiCF(1, 1, 1)	0.25	**77.18**	**60.33**	**69.35**	**71.31**
SiCF(10, 1, 1)	0.25	76.71	60.07	68.73	71.11
SiCF(0.1, 1, 1)	0.25	76.67	59.80	68.77	70.80
SiCF(1, 10, 1)	0.25	76.42	59.42	68.42	70.44
SiCF(1, 0.1, 1)	0.25	76.96	59.89	68.48	70.81
SiCF(1, 1, 10)	0.25	76.11	58.41	67.23	69.43
SiCF(1, 1, 0.1)	0.25	76.24	59.27	67.95	70.41

Table 22: Parameter analysis of uncertainty estimation results on TODSUM 2:90. The values in the brackets are α, β, and γ, respectively.

	ROUGE-1		ROUGE-2		ROUGE-L		BERTScore-F	
	0-50% mean	0-90% mean	0-50% mean	0-90% mean	0-50% mean	0-90% mean	0-50% mean	0-90% mean
SiCF(1, 1, 1)	83.08	**87.92**	70.61	**78.90**	77.38	**83.89**	78.85	**84.91**
SiCF(10, 1, 1)	**83.16**	87.89	**70.73**	78.87	**77.50**	83.85	**78.98**	**84.91**
SiCF(0.1, 1, 1)	82.85	87.62	70.19	78.45	76.90	83.38	78.42	84.45
SiCF(1, 10, 1)	82.70	87.46	70.14	78.33	77.00	83.43	78.48	84.47
SiCF(1, 0.1, 1)	83.14	**87.92**	70.60	78.83	77.20	83.66	78.73	84.75
SiCF(1, 1, 10)	82.80	87.50	69.97	78.14	76.45	82.86	78.07	84.03
SiCF(1, 1, 0.1)	82.90	87.77	70.44	78.76	77.31	83.84	78.77	84.87